Psychological Bulletin

HARRY HELSON, Editor
University of Texas

CONTENTS

Incentive Magnitude, Learning, and Performance in Animals..... BENJAMIN H. PUBOLS, JR. 89

The Paramorphic Representation of Clinical Judgment..... PAUL J. HOFFMAN 116

The Wechsler Intelligence Scale for Children: Review of a Decade of Research..... WILLIAM M. LITTELL 132

Comments on "Intraclass Correlation vs. Factor Analytic Techniques for Determining Groups of Profiles"..... HAROLD P. BECHTOLDT 157

Reply to Professor Bechtoldt's Critique..... ERNEST A. HAGGARD 163

A Review of Hearing in Amphibians and Reptiles..... THOMAS E. MCGILL 165

Published Bimonthly by the
American Psychological Association

VOL. 57, NO. 2 MARCH, 1960

Consulting Editors

R. R. BLAKE
University of Texas

W. R. GARNER
Johns Hopkins University

J. P. GUILFORD
University of Southern California

W. H. HOLTZMAN
University of Texas

D. W. MACKINNON
University of California, Berkeley

L. J. POSTMAN
University of California, Berkeley

S. B. SELLS
Texas Christian University

W. A. WILSON, JR.
University of Colorado

The *Psychological Bulletin* contains evaluative reviews of research literature and articles on research methodology in psychology. This JOURNAL *does not publish reports of original research or original theoretical articles.*

Manuscripts should be sent to the Editor, Harry Helson, Department of Psychology, University of California, Berkeley 4, California.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the *Publication Manual of the American Psychological Association* (1957 Revision). Special attention should be given to the section on the preparation of the references (pp. 50–60), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted in *duplicate*, one of which should be an original typed copy; author's name should appear only on title page. Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail and to check carefully the typing of the final copy.

Reprints. Fifty free reprints are given to contributors of articles and notes.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

SARAH WOMACK
Editorial Assistant

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D. C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Menasha, Wisconsin
and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second class matter at Washington, D.C., and Menasha, Wisconsin. Second class postage paid at Washington, D.C., and at additional mailing offices. Printed in U.S.A.

Copyright, 1960, by The American Psychological Association, Inc.

VOL. 57, NO. 2 MARCH, 1960

Psychological Bulletin

INCENTIVE MAGNITUDE, LEARNING, AND PERFORMANCE IN ANIMALS[1]

BENJAMIN H. PUBOLS, JR.[2]
University of Miami

Interest in the effects of variations in incentive magnitude on learning and performance in animals has increased markedly during the last decade, and there are now more than 75 experimental papers dealing with various aspects of this topic. These papers have been concerned, for the most part, with the following problems:

1. What is the effect of incentive magnitude on rate of learning and asymptotic performance?
2. Does variation in incentive magnitude affect learning, performance, or both?
3. What is the effect of incentive magnitude on resistance to experimental extinction?
4. Does incentive magnitude interact with any other parameters of reinforcement, and if so, in what way?
5. What is the nature of the mechanism, or dimension of reinforcement, whose quantitative variation produces behavioral changes?
6. Are there any differences in performance in terms of whether a given *S* experiences only one incentive magnitude in a given situation, or more than one allowing possible comparisons?
7. What is the effect of incentive magnitude on acquired reward value?

This paper will follow an outline suggested by these questions, and in each case an attempt will be made to draw conclusions as definite as the available evidence allows.

A brief review of some general methodological considerations would seem to be in order to help provide a framework for the discussions which will follow. A variety of types of experimental operation have been used by different investigators in the manipulation of incentive magnitude. These operations have tended to fall into one of five classes, usually involving the manipulation of food in some form, but sometimes water, or a receptive female of the same species.

These classes of operations are: (*a*) variation in weight, volume, or size of a single incentive unit; (*b*) variation in number of equal-weight incentive units; (*c*) variation in duration of exposure to the incentive; (*d*) variation in concentration of sugar, usually sucrose, solutions; and (*e*) variation in incentives which do not alter the effects of deprivation, such as concentration of saccharine, or extent of incomplete sexual behavior.

It should be apparent that the actual dimensions, or mechanisms, of quantitative variation are several, and that they may differ among

[1] The terms incentive, reward, and reinforcement are used interchangeably in this paper, to refer to stimuli which strengthen responses they follow, without implication of subscription to any particular theoretical position. Such phrases as "variation in incentive magnitude," "quantitative variation," and the like, are also used interchangeably. Wisconsin General Test Apparatus is abbreviated WGTA; periodic reinforcement is abbreviated PR.

[2] The author would like to express his appreciation to A. M. Schrier for his critical reading of a preliminary version of this paper.

these classes of operations. Guttman (1953) has pointed out that the expression, "quantity of reinforcing agent," may refer to several dimensions of variation. He lists these as follows: (*a*) amount of nutrient material available for assimilation, in terms of weight or volume; (*b*) stimulation (primarily visual, but may also be olfactory or tactile) derived from the incentive prior to consummatory behavior; (*c*) amount and nature of consummatory activity involved; and (*d*) stimulation from the incentive during consummation (e.g., taste characteristics).

Kling (1956) has further indicated that the third factor, consummatory activity, may be broken down into several subdimensions, which are: number of consummatory responses, duration of consummatory activity, ratio of consummatory to nonconsummatory responses in the goal area, and rate of consummatory responding.

Different combinations of variation result from different operations. Thus, the two classical operations, manipulation of weight and number, each involve simultaneous variation along all four dimensions. Manipulation of duration of incentive exposure typically involves changes in amount of nutrient material and consummatory activity. The manipulation of sugar concentration involves concomitant variation in amount of nutrient material and consummatory stimulation. And the fifth operation may lead to concurrent variation in amount of consummatory activity and stimulation during that activity. The consideration of differences in experimental manipulations will be central to the discussion of mechanisms of reinforcement.

A second category of methodological differences is related to another question raised above. This concerns the number of incentive magnitudes a given *S* experiences. Lawson (1957) calls the method whereby each *S* experiences only one value the "absolute" method, and the method whereby each *S* experiences more than one value the "differential" method. For the most part, studies utilizing the differential method will be considered in a separate section.

To answer the question of whether variations in incentive magnitude affect learning or performance, a special two-phase experimental design will be required. This will be outlined in the appropriate section. However, another device for distinguishing between learning and performance effects, which will be adopted in this paper, requires comment now. This is the use of measures involving time (e.g., latency, running time, speed of response, rate of responding) as measures of performance, and the use of time-independent measures (e.g., errors, trials to criterion) as measures of learning. Although there are perhaps no a priori grounds for making this assignment, the review of the literature which follows should bear out its worth.

It will be noted that certain topics are being omitted which might seem relevant to the issues discussed in this paper. For example, theoretical considerations are minimized. This is not to say that the writer feels them to be unimportant or unsuccessful. Rather, the aim of the present paper is to attempt to order the *empirical evidence* regarding incentive magnitude, so that the theorist will have a clearer picture of the data with which he will work. That is, the paper will attempt an empirical, rather than a theoretical, integration. Theoretical discussions will be found in most of the papers to be reviewed, especially the following: Crespi

(1944); Hull (1943, pp. 124–134; 1952, pp. 140–148); Meyer (1951); Pereboom (1957b); Reynolds (1949); and Spence (1956, pp. 127–148).

Also omitted are latent learning studies of the Blodgett type (e.g., Blodgett, 1929; Tolman & Honzig, 1930). They are especially pertinent to the present paper, as it is possible to interpret them as studies involving a change in incentive magnitude, from a minimal amount to a larger specifiable amount. When so interpreted, their results agree quite well with the results of studies to be reviewed. However, the latent learning experiments have been adequately reviewed elsewhere a number of times (e.g., Thistlethwaite, 1951). Finally, studies in which reinforcement is administered other than peripherally (e.g., intravenously or by stomach fistula) are excluded. In other words, this review will be restricted to studies in which incentives are administered peripherally, and involving exteroceptors.

Rate of Learning and Asymptotic Performance

The evidence to be reported in this section will be based on situations in which animals are given a series of rewarded training trials under one of several quantities of reinforcement, and measurements are made of terminal performance level and rate of approach to this level. Unfortunately, there is not sufficient evidence in all cases that these terminal levels actually represent asymptotic performance. Nevertheless, some measure of performance over the final few training trials is usually given.

Many investigators have taken differences in rate of approach to asymptotic performance level as a function of incentive magnitude to reflect an effect on rate of learning, and different asymptotes to reflect a difference in performance. While the former interpretation may be legitimate, the latter definitely is not. A higher terminal level could be an indication of a greater amount learned, or that amount of learning is the same but performance is superior.

Relevant manipulations have included variations in incentive amount by all five of the operations outlined in the introduction. These will be considered in turn, and, unless otherwise stated, the absolute method of incentive presentation was used.

Variation in Weight and Number of Incentive Units

Apparently the first study of quantitative variation was Grindley's, reported in 1929. He trained five groups of chicks to run down a runway to either 0, 1, 2, 4, or 6 grains of boiled rice. When reciprocal running times on the last five of seven trials were plotted against the number of grains of rice, an increasing, negatively accelerated curve was obtained. It is uncertain just what this curve represents, however. That it represents asymptotic performance with so few trials is questionable.

Wolfe and Kaplon (1941) also used chicks as *S*s. Three groups were run successively on each of three problems, a runway, detour problem, and **T** maze. Running times throughout the 25 or 35 trials given on each problem fell in the following order, shortest to longest, for the three incentive amounts used: four quarter-grains of popcorn, one full grain, and one quarter-grain. Critical ratios on the final training days (consisting of five trials) indicated no significant differences between groups, but several approaching significance in comparisons involving the one quarter-grain group. Inspection of the Wolfe and Kaplon curves indicates a similar

rate of approach to the asymptotes among all groups on all problems.

Crespi (1942) gave runway training to various groups of rats, with the number of incentive units given as reward varied in logarithmic steps. Each incentive unit weighed .02-gm. and the numbers of units used were 1, 4, 16, 64, and 256. Asymptotic running speeds (Trials 21–25 in one experiment, Trials 16–20 in another) were an approximately logarithmic function of the number of incentive units. On the other hand, rate of approach to these asymptotes was approximately constant for the various incentive values. These findings, that the amount of incentive affects performance at asymptote but not rate of approach to the asymptote, have with few exceptions been confirmed in later studies using the absolute method of incentive presentation.

Continuing with the runway studies, Zeaman (1949) varied the weight of single incentive units as follows: .05-, .20-, .40-, .80-, 1.60-, and 2.40-gm. Eighteen or 19 daily trials were given, and equations were fitted describing the decrease in log latency over successive trials, treating Trials 14–19 as asymptotic. Equation constants representing asymptotic performance differed significantly from each other, while those representing rate of approach, or the slope constants, were nearly identical. Conclusions are again clear: Quantitative variation affects terminal level of performance but not rate of approach to this level. A plot of asymptotic log latency as a function of log grams of food yields a decreasing function with slight positive acceleration over the range investigated.

Additional runway studies have corroborated these findings. Among these are the studies of Lawrence and Miller (1947), Metzger, Cotton, and Lewis (1957), and Spence (1956, pp. 130–132).

Two other runway studies require brief comment. Pereboom and Crawford (1958) measured both forward running time and competing response time over 40 trials, and found that incentive magnitude affects both variables, the latter more so. These results suggest that a good deal of the "learning" that is shown by a decrease in running time (or changes in related time measures) may actually reflect the elimination of competing response tendencies, such as exploration, and the like, rather than a marked decrease in *forward* running time.

Gagné (1941) found an increase in rate of acquisition and a decrease in terminal log latency for eight training trials, as a function of incentive amount. However, for purposes of this study, incentive amount was completely confounded with intertrial interval such that the longer rest intervals were associated with larger amounts. The faster learning rate for larger amounts may then be due to the distribution of practice effect, and not to the amount of incentive received.

A study by Hutt (1954) utilizing the bar-pressing response confirms the runway finding that asymptotic performance is positively related to incentive magnitude. Hutt varied both quantity and quality of incentive factorially and assessed their effects on rate of responding under PR. The three quantities were 3-, 12-, and 50-mgm., manipulated by varying the size of the food dipper, and the three qualities were a basic diet plus saccharine (most preferred), basic diet alone, and basic diet plus citric acid (least preferred). The animals

had first received training under continuous reward, and cumulative response curves under PR are essentially linear. Thus it can be concluded that the obtained significant effects of both quantity and quality represent differences in asymptotic performance. The differences were in the expected direction in both cases.

Several studies have employed more complex tasks, such as visual discriminations and **T**-maze problems. Reynolds (1949) trained two groups of rats on a black-white discrimination with a single incentive unit weighing either 30 or 160 mgm. and found that, although mean response times for the two groups differed significantly, differences in trials to criterion were negligible, the means being within one trial of each other. Other investigators (Hopkins, 1955; Schrier, 1956a) have obtained discrimination learning results in substantial agreement with Reynolds.

In another study, Reynolds (1950a) compared acquisition of a simple **T**-maze habit over a constant number of trials with the incentive unit weighing either 30 or 160 mgm. He found a greater percentage of correct responses and faster running times for the larger amount group. Inspection of his curves indicates similar rates of approach but different terminal levels of performance. Coyer (1953) employed a multiple-unit, multiple-choice linear maze and four levels of amount of incentive. Amount failed to have a differential effect on either his learning measure (errors) or his performance measure (running time). Finally, Heyer (1951) employed a five-unit maze and groups of rats under either high or low thirst drive. The high drive *Ss* drank significantly more water in the goal box than did the low drive animals and thus received larger reward. However, the two groups failed to differ in terms of trials or errors to criterion, or times per trial.

Variation in Duration of Incentive Exposure

Kling (1956) employed a runway, the thirst drive, and water as incentive. Amount of water incentive was manipulated factorially with two levels each of duration of exposure to the drinking tube (15- and 120-eec.) and drinking tube diameter (2 and 5 mm.). Several consummatory response measures were obtained and running speeds over the last three of 13 daily trials were found to be unsystematically related to the volume of water consumed per trial, time per trial actually spent drinking, and proportion of goal box time spent drinking. They were positively related, however, to ingestion rate, a measure of consummatory activity. In a similarly designed study, Hellyer (1953) independently varied amount of water reinforcement and drinking tube size, and found that both variables affected runway latencies. There was an inverse correlation between duration of consummatory response and latency.

Fehrer (1956) carried out two experiments, one with a **U** maze, the other with a runway, using the thirst drive and varying both time in the goal box and amount of incentive. Running speeds were found to vary systematically with neither amount of incentive (40-sec. drinking vs. 10-sec. drinking followed by an additional 30 sec. in the now empty goal box) nor time in the goal box (10-sec. drinking vs. 10-sec. drinking followed by 30-sec. delay in the goal box).

Spence (1956) has pointed out that in the majority of the studies of the

type in which weight or number of incentive units are manipulated, duration of time spent in the goal box has been confounded with magnitude, *Ss* typically being allowed to remain in the goal box until, and only until, the incentive has been consumed (the Fehrer study, just described, is an exception to this). Two of his students, Swisher and Czeh, varied magnitude and duration independently by allowing rats with larger amounts time in the goal box equivalent to that of animals with smaller amounts, but then allowing them to finish eating elsewhere. These two studies, one measuring bar-pressing latencies, the other, runway starting speeds (reported on pp. 138–141) both produced evidence that performance varies as a function of duration in the goal box and not amount consumed. Thus, they agree with Kling and Fehrer in the finding that performance does not vary with the actual amount consumed, but disagree with Fehrer's finding that performance did not vary with duration either.

Variation in Concentration of Sugar and Saccharine Solutions

Studies in which variation in incentive magnitude is achieved by manipulating the concentration of sugar or saccharine solutions provide some apparent exceptions to the generalizations that rate of learning is independent of quantity of reinforcing agent, and asymptotic performance is a monotonic function of quantity.

The first investigator to vary concentration of sugar solutions was Guttman (1953). Different groups of rats were presented 4%, 8%, 16%, or 32% sucrose solutions as reinforcement for bar pressing, and the effect was measured on response rate during conditioning, extinction, and PR, and on time for original conditioning. The measures obtained during original conditioning are of immediate concern. Asymptotic response rate, and rate of approach to this asymptote were found to vary with concentration. The relation between rate of approach and concentration was monotonically increasing, but the asymptotic level increased up to 16% and then fell off at 32%. In addition it was found that the time required to emit 500 responses during acquisition was a decreasing function of concentration. The two measures of acquisition rate yield results seemingly inconsistent with findings previously reviewed. The 32% group clearly attained its terminal response rate with fewer reinforcements than did the other groups. But the asymptote for the 32% group is lower than that for either the 8% or the 16% groups. Guttman suggests that this lower asymptote might be due to uncontrolled drinking behavior in competition with bar pressing. If we are to reconcile the finding of faster acquisition rate with increased quantity of incentive in this study, with the results which failed to find such a relationship, one possibility is to give it the status of an artifact. The nonmonotonic asymptote is, as indicated, likely an artifact, and this asymptotic function and the different approach rates are probably interrelated. If the 32% group has less distance to go to reach asymptote, it will get there sooner.

In two studies, Young and Shuford (1954, 1955) investigated the effect of sucrose concentration on a running response. In the first study (1954), latency and running time were measured where concentrations were 2%, 6%, 18%, and 54%. By the end of

18 daily trials the groups had ordered themselves in terms of latencies in the order indicated, greatest to least. But asymptotic running speeds were the same for all groups. Finally, these asymptotic speeds were achieved sooner the higher the concentration, a finding paralleling Guttman's. This finding is possibly artifactual also, because, on the initial trials, *Ss* with higher concentrations were running faster (all *Ss* had had preliminary training with the appropriate concentration, and this can account for the initial gradient), and hence presumably they had less far to go to reach the common asymptote. In the second study (1955), concentration was again varied in logarithmic steps from 2% to 54%, and 25 daily trials were given. Terminal running speeds were nonmonotonically related to concentration.

Hughes (1957) varied both the concentration and volume of saccharine solution, and assessed their effect on latency, running time, and percentage of correct responses in a **T**-maze. For Trials 2–40, the percentage of correct responses increased significantly with both variables. Reciprocal latency and reciprocal running time increased significantly with volume but not concentration. Finally, for the percentage correct response measure, there was a significant interaction between volume and concentration, such that for the smaller of two volumes the effect of concentration was monotonic, but for the larger volume the effect of concentration was nonmonotonic.

Finally, Smith and Duffy (1957) reported faster learning of a **T**-maze habit when the reward was 4 cc. 20% sucrose than when only .1 cc., in terms of increases in percentage correct, and decreases in running time. Here we have another possible exception to the generalization that incentive magnitude does not affect rate of learning. However, their measures of rate of learning were somewhat unorthodox and no direct statistical comparisons in terms of volume were made. In the interests of parsimony, and until this design is repeated, preference is to maintain the original conclusion that learning rate is independent of incentive magnitude.

Zero Incentive Magnitude

A final set of three papers are related in that, in each, one of the magnitudes employed was a zero magnitude. Furchtgott and Rubin (1953) used a two-unit, linear **T** maze and assigned different groups of rats to a single incentive unit weighing either 0, 20, 75, 250, or 2500 mgm. No differences in terms of three measures of rate of learning (trial to criterion, number of *Ss* attaining this criterion, number of errors during the precriterial period) were found except that all groups receiving any positive amount did better than the group receiving zero amount. However, mean running speeds did differentiate the positive amount groups, the two larger reward groups running faster than the two smaller reward groups.

Two other studies support the finding that a positive amount leads to faster learning than zero amount. Seward, Shea, and Elkind (1958) found a significant interaction between incentive amount (a 1-gm. or a .5-gm. pellet vs. no pellet) and trials, indicating a possible effect on rate of learning, but there was no improvement at all when the goal box was empty. Smith and Kinney (1956) found faster bar pressing for a reward of 20% sucrose solution than for 0% (plain water), during a single 26-min. session.

Learning Versus Performance Effects

The tentative conclusions arrived at in the preceding section are that incentive magnitude does not affect rate of learning but that it does affect asymptotic performance. The present section will be concerned with whether this difference in asymptotic performance reflects a difference in amount learned, or in level of performance independent of learning. The experimental differentiation of learning and performance effects requires a two-phase experimental design, comprising a training and a test phase. Training-phase performance preferably should be asymptotic before the test phase is introduced. Measures obtained during the training phase would not, as we have noted, differentiate learning and performance effects, as either could be present, and, if so, their measurements would be confounded. However, where incentive magnitude is varied during the training phase but held constant (or otherwise equated) during the test phase, differences in performance during the test phase would reflect differences in strength of learning. In contrast, where incentive magnitude is held constant during the training phase but varied during the test phase, test-phase performance would reflect effects of contemporary incentive magnitude and hence a direct effect on performance. These interpretations rest on the assumption that learning effects should be relatively permanent and carry over from one phase to the next, while performance effects should not (Lewis & Cotton, 1957; Maher & Wickens, 1954).

An implication of these statements is that, if magnitude affects learning, the approach to the new asymptote under test-phase magnitude (where there has been a change in magnitude in going from one phase to the next) should be gradual, as it was in the training phase. But if magnitude affects momentary performance, the approach to the new asymptote might be either gradual or abrupt. The most compelling evidence that magnitude affects performance independently of learning would come from findings of an abrupt change in performance level. This is not a necessary requirement however, for it is possible that competing response tendencies might prevent sudden performance shifts (see Pereboom: 1957a, 1957b). In contrast, if incentive magnitude affects amount learned, and learning is not construed as a one-trial affair, then the effects of change in magnitude *must* include a gradual change in level of performance.

Variation in Weight and Number of Incentive Units

The earliest study to use training and test phases, as here defined, was Crespi's (1942). In one of his experiments, after rats trained with either one or four incentive units had reached stable asymptotes their magnitudes were both shifted up to 16 units. In another experiment, 64- and 256-unit groups were shifted down to 16 units. In all but the 1-to-16 condition, running speed shifts were abrupt such that after but one exposure to the new incentive amount, the running speed of the animals reached the level of *Ss* trained with 16 units from the start. In the 1-to-16 case this level was reached after two exposures. Additional training at these new values led the animals to "overshoot" in their performance levels. That is, *Ss* with an increase in incentive magnitude ran faster, while those with a decrease ran more slowly, than those trained with 16 units from the start.

Both of these overshooting effects were significantly different from obtained or extrapolated values for the 16-unit rats, and were labeled "elation" and "depression" effects, respectively. Later investigators have dubbed them "positive contrast" and "negative contrast" effects. Elliott (1928) had earlier found an effect similar to the negative contrast effect when rats trained in a multiple **T** maze with bran mash as incentive were shifted to sunflower seed. Crespi's results indicate that the effect of change in incentive magnitude is one of sudden change in performance, and thus suggest that incentive magnitude affects level of performance rather than amount learned.

Zeaman (1949) also studied the effect of sudden shifts in incentive magnitude upon runway behavior. Changes in latency were abrupt, occurring after but one exposure to the new incentive values. In one experiment, the change in latency was directly proportional to the change in amount, and in another, based upon extrapolation, a positive contrast effect was obtained, but a possible negative contrast effect was not significant.

Another study from which information based on abruptness of performance change can be obtained is that of Spence (1956, pp. 130–132). Following 48 runway trials with either .05- or 1.0-gm. incentive, these values were interchanged. Again, changes in response strength were instantaneous. A significant negative contrast effect was obtained, but not a positive contrast effect.

Metzger, Cotton, and Lewis (1957) employed a factorial design in which four groups of rats received either two or eight 45-mgm. pellets during the first and second 10 runway trials. Speed of running during the second (test-) phase varied only as a function of incentive magnitude during that phase. In assessing contrast effects, this study has the obvious advantage over the others that comparisons need not be made on the basis of extrapolation to what performance might have been had incentive shifts not occurred, for, according to the design, only half of the *Ss* experienced a change in incentive amount. Although changes in running time occurred within one or two trials following incentive shift, there was no evidence for either positive or negative contrast effects.

It is clear from these studies that quantitative variation in incentive affects level of performance rather than amount learned. But how are the contrast effects to be interpreted? Crespi and Zeaman gave their *Ss* 18–20 training trials before shifting incentive values; Spence gave his *Ss* 48 preshift trials. Both Crespi and Zeaman found positive contrast effects but Spence did not. Spence argues that those found by Crespi and Zeaman were artifacts, due to the possibility that their rats had not had enough trials to reach a true asymptote before the shifts were introduced. His evidence, reported above, supports this interpretation. In other words, perhaps performance would have continued to improve anyway, without the increase in incentive magnitude. However, that Metzger et al. gave only 10 preshift trials, fewer than any of the other experimenters, and failed to demonstrate positive contrast effects, argues against Spence's position. Out of four studies (excluding the Elliott study) only two demonstrated each type of effect, and the one study which maintained control *Ss* not shifted yielded negative results for both types of effect. Before the matter is resolved more information is needed on the conditions under which

contrast effects would and would not be expected to appear. Pereboom (1957b) has suggested that runway familiarity and competing exploratory behavior are important variables in this connection. Two related possibilities suggest themselves. First it should be noted that both Crespi and Metzger et al. employed measures based on running time and that the runway used by Crespi, who obtained positive results, was five times as long as the one used by Metzger et al., who obtained negative results. Alley length may be an important variable. Second, regarding the extent of preshift training, if animals are given sufficient trials to bring them to the true "physiological limit," it would be difficult if not impossible for low-to-high *S*s to exceed in performance animals already on the higher amount.

One study involving the training-test paradigm used a learning situation more complex than the runway. Maher and Wickens (1954) used a test situation requiring relearning under a drive different from the one used in training. Their rats were given 20 trials in a 14-unit multiple T maze with 22-hr. food deprivation and either one or five pellets as incentive. No differences in rate of acquisition, in terms of errors, were found, but terminal speeds (Trials 19–20) were greater in the five pellet group, findings consistent with those reviewed in the preceding section. Following a lapse of 2½ months, the maze was relearned under 22-hrs. water deprivation and a constant amount of water reward. During this relearning test, neither time nor errors, as a function of training-phase magnitude, were significantly different.

Variation in Concentration of Sugar Solutions

Following conditioning, extinction, and reconditioning, Guttman's (1953) rats were tested under 1-min. PR for five daily ½-hr. sessions. Rate of responding was an increasing logarithmic function of concentration. Note that in this study, the test phase involved, not a shift in incentive magnitude, but a shift from continuous reinforcement to PR, and that the relation between magnitude and performance changed from nonmonotonic to monotonic.

Young and Shuford (1954) shifted the concentrations of their rats following the 18 conditioning trials reported above. The general effect was a decrease in total time (latency plus running time) when concentration was increased and an increase in total time when concentration was decreased. The former effect was more pronounced than the latter and this difference was attributed to a practice effect.

Dufort and Kimble (1956) gave rats 20 training trials in a runway with a 10% sugar solution in one of five bottle caps located on the goal platform. Following this, they were split and continued under either 0% (extinction), 5%, 10%, or 20% concentration for an additional 40 trials. Following the shift in incentive concentration, the 20% group changed its percentage correct responses abruptly, while the other groups changed more gradually. The 20% and 10% groups continued to improve (indicating in the latter case that asymptotic performance had not been achieved before differential training was introduced), both reaching 100% correct by the end of training, while the 5% and 0% groups declined in percentage correct. Differences in reciprocal running times were also assessed, and again the change in performance of the 20% group was abrupt while the other changes were gradual. Terminal running speeds were a positive function of concentration, and the rate of ap-

proach to these speeds was the same in all cases except for the 20% group, whose slope constant was greater than that of the other groups. This difference may be due to the fact that asymptotic performance had not been achieved before differential concentrations were introduced. Dufort and Kimble conclude, " . . . the indication is that the effect of changing the amount of reinforcement is at least partly on performance rather than habit" (p. 190). We would conclude that the effect is exclusively on performance. They also suggest, on the basis of the near identity of the 5% and 0% slopes, that extinction be conceptualized as the limiting case of decreasing the amount of incentive.

Resistance to Extinction

Studies of the effect of incentive magnitude on resistance to extinction might profitably be grouped in terms of the extinction measure employed. Some investigators have used as their measure trials to an extinction criterion, while others have assessed performance over a constant number of extinction trials. One report, however, does not give enough information for this classification. This was the study by Fitts (1940), who compared the extinction performance of rats following the 10 experimental conditions of 1, 5, 10, 20, or 30 rewarded bar-pressing responses with either .2- or 10.0-gm. incentive. For all numbers of rewards, resistance to extinction was greater following the larger of the two amounts.

Trials to an Extinction Criterion.

Three studies which employed a measure of trials to a specified extinction criterion found unsystematic effects of prior incentive magnitude, while one found a systematic effect. Thus, Lawrence and Miller (1947) found an insignificant difference in trials to running response extinction criteria of either 3- or 5-min. latency as a function of prior amount (one or four pellets). Reynolds (1950b) administered rewards to three groups of rats for 25 consecutive bar-pressing responses. The three groups received one 60-mgm. pellet, two such pellets, or one 160-mgm. pellet for each response. No significant differences were obtained between any of the groups in terms of responses to a no-response criterion of 5 min. In a related study, Reynolds, Marx, and Henderson (1952) obtained no significant differences in trials to extinction of a bar-pressing response following 120 vs. 30 mgm. reward.

Young and Shuford (1955) extinguished their *Ss* with distilled water following the 25 training trials previously mentioned. The slope of the extinction curves of running speeds varied directly with prior concentration. Nevertheless, *Ss* with lower concentrations reached an extinction criterion sooner than those with higher concentrations, perhaps because they had less far to go to reach this criterion.

Performance over a Constant Number of Extinction Trials

Zeaman (1949) subjected the *Ss* of his various experiments to extinction of the running response after they had completed their rewarded training, either with or without test-phases intervening between training and extinction. In all three of his experiments, the effect of previous magnitude was to alter the rate of approach to a final common performance level. Gagné (1941) reported a higher terminal extinction level (after only five extinction trials) for greater reward amounts, confounded with longer intertrial intervals. Metzger et al. (1957) extinguished their *Ss* after the training and test phases considered earlier. Using running times

during Extinction Trials 2–11, or Extinction Trial 2 alone, training-phase magnitude had no effect. However, test-phase magnitude had a significant effect on both measures, greater resistance following the larger amount.

Following PR under three levels each of quantity and quality, Hutt's (1954) rats were given two ½-hr. extinction sessions. Responses during extinction varied as a function of both variables, more responses being emitted following PR with larger amounts and preferred qualities. Guttman (1953) found rate of responding during an initial 5-min. of extinction to be an increasing monotonic function of the concentration of sucrose used during conditioning.

Fehrer (1956) found greater resistance to extinction (in terms of both running speeds over a constant number of trials, and trials to various criteria) following the condition of 10-sec. drinking followed by 30-sec. postreinforcement delay in the goal box, than following either 10-sec. or 40-sec. drinking time in the goal box.

According to the assumption that learning effects are relatively permanent while performance effects are momentary, we would be led to conclude, on the basis of extinction performance differences found for a constant number of trials, that magnitude of reward does affect learning when the absolute method is used. However, it might be that the higher performance level during extinction following larger amounts is due to the fact that terminal acquisition level is higher the greater the amount of incentive. In fact this appears to be the case. Metzger et al. (1957) reported results of an analysis of covariance upon their extinction scores, where performance levels at the beginning of extinction were equated in terms of running times during the last five reinforced trials. Under these circumstances, the effect of test-phase amount on extinction disappeared. They conclude, " . . . reward affects performance on extinction through differential levels of performance just prior to extinction rather than affecting performance on extinction directly" (p. 188). If we apply this interpretation to the other studies in which performance was measured over a constant number of extinction trials, the conclusion that incentive magnitude does not affect amount of learning is maintained. This interpretation might also account for the apparently anomolous results of Young and Shuford (1955) and Fehrer (1956) on trials to extinction criterion.

Interaction of Incentive Magnitude with Other Variables

Several investigators have manipulated incentive magnitude factorially with other variables. While their results agree with previous conclusions, they also provide information on possible interactions of this variable with other variables. In particular, studies have allowed the assessment of interactions of incentive magnitude with quality of reward, drive level, and partial reinforcement.

Quality of Reward

Hutt (1954) found that the effects of quantity and quality of reward on rate of bar pressing during both PR and extinction were independent of each other, the interaction F ratio in the former case being less than one.

Drive Level

Seward, Shea, and Elkind (1958) factorially varied incentive magni-

tude and length of food deprivation, and assessed their effect on running speed. Both main effects were significant, as was the interaction between them. But this interaction is based on comparisons involving zero reward amount and zero hours deprivation. No learning took place under satiation or when the goal box was empty. That the interaction might represent a special case found only when values of either incentive magnitude or length of deprivation are zero, as suggested by Seward et al., is attested by a recent study by Reynolds and Pavlik (1958). They varied incentive magnitude (.1, 1.0, and 2.0 gm.) and deprivation time (3, 22, and 44 hours) factorially over 72 runway trials and reported reciprocal latencies over the last 20 of these trials. Differences as a function of both incentive amount and deprivation were in the expected direction and significant. However, the interaction between these two variables was not significant ($F < 1.00$).

Reynolds et al. (1952), in contrast, reported a significant interaction between amount of incentive and drive level on trials to an extinction criterion of 5-min. with no bar-pressing, such that high drive–high reward and low drive–low reward animals extinguished more rapidly than high drive–low reward or low drive–high reward animals. Reynolds and Pavlik suggest that this difference might be a function of the different response measures employed in the two studies. Another possibility lies in the definition of drive. For Reynolds and Pavlik, drive was defined in terms of the length of food deprivation. In contrast, Reynolds et al. manipulated both amount fed and deprivation time simultaneously. In any event, it would seem that more research is needed to specify the conditions of both drive and reward under which an interaction between them would be expected.

Partial Reinforcement

Hulse (1958) found a greater running speed over the last nine of 25 daily runway trials when the incentive was a 1.0 gm. food pellet than when it was a .08 gm. pellet, and when reinforcement was continuous than when appearing on only 46% of the trials. He also measured running speeds over a constant number of extinction trials and found a significant interaction between the effects of prior percentage of reinforcement and amount of reinforcement such that, with partial reinforcement, resistance to extinction was greater following a larger amount, but with continuous reinforcement, the reverse effect obtained.

Mechanisms of Reinforcement

In the introduction, four possible mechanisms of quantitative variation were outlined: amount of nutrient material available for assimilation, preconsummatory stimulation, consummatory activity, and consummatory stimulation. It was further indicated that different experimental operations produce different combinations of variation. Thus, decisions regarding mechanisms of reinforcement must involve comparisons of experiments utilizing the different operations.

We are actually concerned here with the conditions of reinforcement, and a distinction must be made between *necessary* and *sufficient* conditions. If quantitative variation is shown to affect behavior when one of the mechanisms is held constant, this would indicate that variation in that mechanism is not necessary for

incentive magnitude effects. On the other hand, lack of behavioral variation would indicate that that mechanism is necessary. If only a single mechanism is varied and this produces behavioral variations, it may be concluded that variation in that mechanism is sufficient. But if there is no behavioral variation, it must be that that mechanism is not sufficient, or that a necessary variation (if there are any) is lacking. Conclusions regarding necessity and sufficiency must be reached through comparison of studies and a process of elimination. Because experimental manipulations of weight and number confound all four possible mechanisms, studies of this sort are not crucial to the problem.

Preconsummatory Stimulation

That variations in preconsummatory stimulation are not necessary is attested by the results of studies in which either duration of incentive exposure, or concentration of sugar or saccharine are manipulated. In all of these cases, preconsummatory stimulation is presumably constant, and yet behavioral variations, previously reviewed, have been found.

Two studies in which preconsummatory visual stimulation was independently manipulated indicate that variation in this mechanism is not sufficient either. McKelvey (1956) factorially varied the duration of incentive exposure (30 vs. 180 sec.) and visual size of the incentive (food tray of 1-in. diameter vs. 2-in. diameter), during acquisition of a black-white discrimination. Duration of reward affected the performance measure, running time, but not the acquisition measure, errors. But by neither measure was preconsummatory stimulation shown to have any systematic effect on behavior.

This finding is corroborated in a study of delayed response performance by chimpanzees using a modified differential method, by Cowles and Nissen (1937). In one condition of their experiment, the chimps were shown either a large piece of orange without skin or a small piece with skin (the skin differences were used to equate eating times), but after the delay they always found the smaller. In the other condition, the chimps found the size they had been shown prior to the delay. Thus, with the first procedure, preconsummatory stimulation is varied but the amount remains constant; with the second procedure the two factors are confounded. With the former procedure, percentage of correct responses, latencies, and response times did not differ systematically in terms of whether the pre delay stimulus was large or small, but with the latter procedure, differences were in the expected direction and, for the most part, statistically significant.

Amount of Nutrient Material

A number of studies involving the use of, and variations in the magnitude of, incentives which do not alter the effects of deprivation indicate that variation in the amount of nutrient material available for assimilation is not a necessary condition of reinforcement. Amount of nutrient material is effectively controlled in these studies, by elimination.

The pioneer study with saccharine incentives is that of Sheffield and Roby (1950). This was essentially a demonstrational study, designed to determine if saccharine could act effectively as a reinforcer. Although three experiments were conducted, only the third will be described here. For this experiment, rats learned a position habit in a **T** maze. Food-deprived *Ss* were given 42 trials with saccharine solution in one arm of the

maze and tap water in the other. There developed a significant increase in choices of the saccharine side, as well as an increase in rate of ingestion and decrease in running time.

The study of Hughes (1957), discussed in an earlier section, indicates that, not only can rats learn in the absence of reinforcement by nutrient material, but that performance is differentially affected by saccharine concentration.

Two studies using sexually motivated male rats as *Ss* with receptive females as incentive also indicate that behavior may be modified in the absence of alterations in the effects of deprivation. Sheffield, Wulff, and Backer (1951) compared running speeds when the incentive was either a female rat in heat or another male. Even though ejaculation was not permitted, the *Ss* ran faster to the receptive female than to the male incentive. Kagan (1955) found a greater percentage of correct responses and faster running speed in a **T** maze when the incentive was copulation with ejaculation than when only intromission was permitted. Performance under both of these conditions was in turn superior to that when only mounting was allowed.

Several studies, already discussed, indicate that, when incentives are administered peripherally, variation in amount of nutrient material is not a sufficient mechanism of reinforcement. Kling (1956) found running speeds to be unrelated to volume of water consumed per trial, and Fehrer (1956) also found running speeds to be independent of amount of water reinforcement. Swisher and Czeh (Spence, 1956) found performance to be unrelated to amount of food consumed.

It should be emphasized that the conclusion that nutrient material, and hence that variations in amount of nutrient material is not a sufficient condition, applies only to situations in which incentives are administered peripherally. If we include evidence from experiments in which peripheral factors are by-passed, then the conclusion must be that variation in nutrient material is a sufficient condition. See for example, the studies of Coppock and Chambers (1954), and Miller and Kessen (1952). But the remaining evidence will indicate that, when incentives are peripherally administered, the reinforcing mechanism is itself peripheral.

Amount of Consummatory Activity

When incentive magnitude is varied by manipulating the concentration of sucrose solutions, the amount of consummatory activity is presumably held constant. The studies reviewed earlier indicate that performance does vary as a function of sucrose concentration, so it would seem that variation in amount of consummatory activity is not a necessary mechanism.

However, variation in consummatory activity does seem to be a sufficient condition. A number of studies in which ingestion rate was measured would so indicate. For example, Sheffield, Roby, and Campbell (1954) found a high positive correlation between running speed and ingestion rate (incentives were water, saccharine, dextrose, or dextrose plus saccharine solutions), direct evidence that performance is an increasing function of amount of consummatory activity. The Kling (1956) study also indicates that the important consummatory activity variable is ingestion rate.

Supporting evidence comes from the finding of Sheffield, Wulff, and Backer (1951) of a positive relationship between running speed and per-

centage of opportunities to attempt copulation during which the attempt was actually made.

Amount of Consummatory Stimulation

When duration of incentive exposure is manipulated, the nature of the consummatory stimulation is held constant (although its duration may vary). Several studies of this type (e.g., Kling, 1956; Spence, 1956) have found performance variations, suggesting that variations in consummatory stimulation are not necessary.

That they are sufficient, however, is indicated in a study by Cockrell (1952), similar in design to Guttman's (1953), but manipulating concentration of saccharine rather than of sucrose solutions. By his procedure, the only mechanism assumed to vary was consummatory stimulation (taste). During conditioning under continuous reinforcement, barpressing rates reached nonmonotonic asymptotes as a function of concentration. During extinction and PR, rate of responding was approximately a linear function of the logarithm of saccharine concentration.

We may close this section with the tentative conclusion that none of the mechanisms of reinforcement are necessary for bringing out performance differences as a function of incentive magnitude, but that variations in either amount of consummatory activity or stimulation associated with that activity are sufficient conditions. By way of qualification, it should be pointed out that there is some difficulty in manipulating consummatory stimulation while holding consummatory activity constant (cf. Guttman, 1953). Thus it may be that only the *concurrent* variation in consummatory activity and consummatory stimulation will prove to be sufficient.

The Comparison of Incentive Magnitudes: The Differential Method

The term *differential*, as used here, will be given a somewhat broader meaning than that originally assigned by Lawson (1957). Several differential procedures may be distinguished:

1. Single Problem Method—Throughout the experiment there is but a single task or problem before the *S*, but in different stages of the experiment he receives different amounts, for example, according to a latin square design.

2. Simultaneous Discrimination—Selective learning is involved such that one of two responses is followed by a greater amount, the other by a lesser amount. *S* must learn to discriminate magnitudes.

3. Successive Discrimination—On successive trials, *S* makes one response followed by a larger amount, or a second response followed by a smaller amount.

4. Successive Problems—A series of problems are to be learned, each problem correlated with a different incentive amount. The "learning sets" paradigm (Harlow, 1949) provides a familiar example.

It will be noted that in some cases the distinction between absolute and differential methods will appear arbitrary. Certainly the previously discussed training-test-phase design under the absolute method involves exposure by the same *S* to more than one incentive amount. However, separate measurements were made under each phase of the experiment, and statistical analyses were separate. In contrast, studies to be reviewed in this section are ones in which a single analysis is performed on data obtained under variable incentive conditions for the same *S*.

Single Problem Method

Gantt (1938; also reported in Hull, 1943, p. 125) presented a curve of asymptotic (?) conditioned salivation in a single dog as a function of incentive magnitude—½, 1, 2, and 12 grams of food. The curve represents an increasing, negatively accelerated function. It is interesting to note that this is apparently the only study of incentive magnitude in the literature based upon the classical conditioning procedure.

Nissen and Elder (1935) varied both the size of a single piece of banana and the number of constant-weight pieces and assessed the effect on the limits of delayed response at an accuracy of 80% or better. The total number of grams varied from 3 to 20, and increases in delay limits accompanied increases in incentive amount, and the reverse. A perseveration effect—the opposite of contrast effects—was also noted. A second delayed response study (Cowles & Nissen, 1937) has previously been cited. They found fewer errors and shorter latencies with large than with small incentives.

Fletcher (1940) assessed the effects of incentive magnitude (length of banana slice) on the performance of pulling responses by chimpanzees and found a positive relationship between frequency of response and incentive length. In addition, various time measures were significantly affected, among them, response latency, pulling time, and the number of tugs per trial.

At this point, Tinklepaugh's early study (1928) might be mentioned. He noted in reward substitution experiments that when monkeys observed the experimenter place two pieces of food under a cup, but later found only one piece, certain emotional and searching responses ensued. This behavior did not occur when one small piece was substituted for one large piece.

Michels (1957) used a latin square design to assess the effects of amount of reinforcement (.5, 1, 2, or 4 peanuts) on latency of response to a single test object. The WGTA was used and the rhesus monkeys who served as *S*s were on a 20% partial reinforcement schedule throughout. Latencies with the .5-peanut reward were significantly greater than with the three larger rewards, which did not differ reliably from each other.

The remaining studies utilizing the single problem method are all operant conditioning studies. Jenkins and Clayton (1949) allowed a group of pigeons either 2-sec. or 5-sec. exposure time to the food magazine in a counterbalanced order during PR and found a faster rate of key-pecking with the longer duration. Roughly twice as many consummatory responses were emitted during the 5-sec. exposure than during the 2-sec. exposure.

Following the training previously described, 20 of Guttman's (1953) rats were given further PR training, each *S* under each of the four sucrose concentrations used. A linear relationship was found between rate of responding and the logarithm of concentration, as had been found with independent groups under PR. In a second study, Guttman (1954) compared the reward values of various sucrose and glucose concentrations, varied in equal logarithmic steps from 2% to 32%. For both sucrose and glucose, the relationship between rate of bar pressing under an aperiodic schedule, and log concentration, was linear. Throughout, rate was higher with sucrose than with the corresponding glucose value. Two latin squares were used such

that any *S* experienced all concentrations of one substance but not the other.

Conrad and Sidman (1956) investigated the effect of sucrose solution concentration on bar pressing by rhesus monkeys. Each *S* experienced each of seven concentrations, ranging from 0% to 60%. The maximum rate occurred between 15% and 30%, with a decline at 60%. Verhave (1956) assessed the effects of seven different sucrose concentrations presented to the same rats on response latency and rate for a two-member chain (pulling responses). Concentrations varied from 2% to 32%. Response rate increased and latency decreased as a function of concentration, both functions reaching their asymptote in the neighborhood of 20–21% concentration.

Collier and Siskel (1959) and Collier (1958) have attempted an examination of the factors producing nonmonotonic incentive functions, which have been found with either the absolute or the differential method. Collier and Siskel varied sucrose concentration (4%, 8%, 16%, and 32%) and PR interval (.5-, 1-, 2-, and 4-min.) factorially and assessed their effects on bar-pressing rate in rats. Both main effects, as well as their interaction, were significant. The nature of the interaction was such that the nonmonotonicity of the obtained concentration function was itself a decreasing function of the inter-reinforcement interval. Collier (1958) discusses these and other data and concludes that a nonmonotonic relationship will most likely be found under any of the following conditions: an extended test-session, large volume (Hughes' [1957] study verifies this assertion), high concentration, or short time between reinforcements.

Simultaneous Discrimination

Festinger (1943) allowed rats 10-sec. exposure to whole wheat grain in one arm of a discrimination box and 1-sec. exposure in the other arm. By the end of 96 trials, the rats were responding on free-choice trials to the side with the greater exposure time at an above-chance level. Denny and King (1955) replicated the study with .7-gm. reward in one goal box of a **T** maze and .1-gm. in the other. By the end of 84 training trials, a preference for the larger reward side had developed. Discrimination training was followed by reversal training, with the large- and small-reward sides interchanged. By the end of 72 reversal trials the *S*s were running to the opposite side with a frequency greater than chance.

Pereboom (1957a) gave rats two free- and two forced-choice trials per day in a **T** maze for either 10 or 21 days. Five food pellets were found on one side, one on the other. Following this training, the large and small rewards had their sides interchanged until a total of 35 days had been completed. As a third stage of the experiment, for an additional 12 days the reward amounts were equalized on the two sides. By the end of the initial 10 training days, all *S*s were responding consistently to the large reward side. Reversal, as in the Denny and King study, was gradual, but complete by the end of reversal training. During the subsequent reward-equalization phase, the *S*s tended toward chance performance. The gradual reversal found by both Denny and King, and Pereboom, is to be contrasted with the abrupt performance changes found when incentive magnitudes are altered in the runway. Pereboom interprets this inconsistency in terms of competing exploratory behavior which retards

reversal learning, and would be expected to be greater in the more complex **T** maze than in the runway.

Wike and Barrientos (1957) held amount of nutrient material and consummatory stimulation constant while allowing only duration of consummatory activity to vary. This was achieved by varying the diameter of a drinking tube from which the water incentive (for thirsty rats) was to be obtained. With the smaller diameter more consummatory activity was required to attain a constant amount of water. The two diameters were pitted against each other in opposite arms of a **T** maze and the rats were required to learn a position discrimination on this basis. By the end of the 27 daily trials given, 85% of responses were to the side with the smaller diameter, a level significantly above chance. These results nicely confirm an earlier conclusion that variations in neither amount of nutrient material nor consummatory stimulation are necessary, but variation in consummatory activity is a sufficient mechanism.

Successive Discrimination

D'Amato (1955) gave rats training on a successive discrimination problem in a runway such that on half the trials a larger incentive was associated with a goal box of one color while on the other half a smaller incentive was associated with a differently colored goal box. By the end of 70 training trials, the *Ss* were running significantly faster for the larger reward than for the smaller.

Greene (1953) employed a design utilizing successive discrimination during a training-phase and simultaneous discrimination during a test-phase. By a factorial design it was possible to assess learning vs. performance effects. During the training-phase, rats found either a large pellet in a black goal box and a small pellet in a white one, or the reverse. During the test-phase, either a large or a small reward was found in the black arm of the apparatus, the white arm being empty. Number of errors and number of trials to criterion during the test-phase were significantly less for the animals having the larger reward associated with black during the training-phase, but performance did not vary systematically as a function of contemporary amount. This result is flatly counter to test-phase results reported earlier using the absolute method. Here we have evidence for a persistent effect of training-phase amount on later performance, which can be interpreted to mean that, with the differential method, amount of incentive affects learning. This conclusion will be verified in subsequently discussed papers.

Powell and Perkins (1957) manipulated duration of incentive exposure during successive discrimination training and found it to influence simultaneous discrimination test-phase performance, confirming Greene's results.

Successive Problems

Experiments in which *Ss* are trained on a number of consecutive problems, with different incentive magnitudes associated with different problems, yield clear-cut evidence that incentive magnitude affects learning.

The first, and probably the most significant, of these experiments was published by Meyer in 1951. Meyer trained eight rhesus monkeys, highly sophisticated on several tasks including discrimination reversal, on a series of 64 different discrimination

reversal problems in the WGTA. Each problem used a single pair of objects throughout an original learning phase and four reversals (ABABA). Response to the correct object yielded a constant amount of incentive for any given reversal but varying amounts over the four reversals of that problem. Statistical analysis was of reversal errors as a function of the following factors: practice, or number of prior problems; reversal number within a problem; prereversal reward (one or three pieces of raisin or peanut); postreversal reward (one or three incentive units); prereversal criterion (two or four successive correct responses); and postreversal criterion (same). The critical comparisons for the present context are those involving pre- and postreversal amounts, and stage of practice.

Both pre- and postreversal amount affected reversal performance, and the first-order interactions between pre- and postreversal amount, and between postreversal amount and practice, were significant. What this means is that both prereversal amount (training-phase amount) and postreversal amount (test-phase amount), affect postreversal performance. The learning of discrimination reversal per se may be ruled out as a factor in that *Ss* entered this experiment with a considerable amount of reversal learning experience, and over-all changes in performance with stage of practice were insignificant (total errors on the first half of the experiment were 1,402, on the second half, 1,404). So it would seem that we have unambiguous evidence that incentive amount affects learning as well as performance independently of learning. The significant interactions imply that the effects of pre- and postreversal amounts are not independent of each other. However, this nonindependence was apparent on early problems only. By the termination of the experiment, the four treatment combinations of pre- and postreversal amounts were ordered in their effect, most to least errors, as follows: 3-1, 1-1, 3-3, and 1-3. This ordering indicates possible long-lasting contrast effects, both positive and negative, that are slow to develop. The measure Meyer employed was errors to criterion, a measure we have taken to represent effects on learning. Thus, the results of the experiment taken as a whole suggest the following conclusion: Incentive magnitude affects learning when *Ss* experience different incentive magnitudes, but the *Ss* must first learn to discriminate magnitudes. Once magnitude discrimination is achieved, differential effects on problem learning emerge.

This conclusion receives support from a more recent study by Schrier and Harlow (1956). They used color discrimination problems to assess the effects of incentive amount (one, two, or four 2.2-gm. pellets), problem difficulty, and practice, on a learning measure, percentage of correct responses, in Java monkeys. All main effects were significant, as was the practice by amount interaction. This reflects the finding that performance begins at about the same level for all amounts, but subsequently the learning-set curves diverge, so that different asymptotes are approached. Schrier and Harlow say, " . . . that a learning process is involved which influences the perception of, and in turn, the response to, varied amounts of incentive, and that such learning is independent of discrimination learning per se" (p. 120).

Meyer and Harlow (1952) investigated the effects of incentive amount (1, 2, 3, or 4 units) on delayed response performance of rhesus mon-

keys, using the WGTA. Percentage of errors decreased with increasing incentive amounts, and the incentive function changed with practice from one with slight positive acceleration to negative acceleration.

Davis (1956) reported results of two experiments on problem solving in monkeys. He found that an increase in incentive amount led to an increase in balks (i.e., decrease in percentage of responses), but no change in errors, on visual size discrimination problems. The unexpected effect on balks was apparently due to satiation effects with the larger amounts. In a related study, a reduced-cue problem, three raisins produced fewer Trial 2 errors than 1, but there were no significant differences in the number of balks.

Finally, Leary (1958) compared four conditions of reward in their effect on serial discrimination learning in rhesus monkeys. The four conditions, orthogonal to four 10-pair lists in a greco-latin square, were: two peanuts for all pairs of a list; one-half peanut for all pairs; two peanuts for five pairs, one-half for the other five; and two peanuts for two pairs, one-half peanut for the remaining eight. The two homogeneous reward conditions led to performance superior to that with heterogeneous reward, the difference attributable primarily to differences in errors on small reward pairs under the two conditions.

Paired-Comparisons Studies

Two studies, although employing variations on the differential method, do not fit into any of the above categories. These are paired-comparisons studies of food preference as a function of magnitude. Harlow and Meyer (1952) gave seven naive rhesus monkeys 400 WGTA trials in which they were to choose between two peanut amounts: $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, or 4 peanuts. This gave 10 pairs. Resulting percentage choice scores were converted to scale values by a method similar to Thurstone's Case III, resulting in an increasing linear relationship between scale value and log incentive amount. Fay, Miller, and Harlow (1953) had their monkeys choose between a preferred quality (peanut or bread) and a nonpreferred quality (potato), when each of five values of the preferred was paired with each of two values of the nonpreferred. They found that, as the ratio of preferred to nonpreferred amount increased (from 1/64 to 1/2), the choices of the preferred food increased. Further, the percentage choice depended, not upon absolute amounts compared, but upon their ratios.

Comparison of Absolute and Differential Methods

Several studies afford a more direct comparison of the absolute and differential methods than has been possible in studies reviewed so far, in that the two methods are compared within a single experimental design.

Logan, Beier, and Ellis (1955) compared runway speeds during acquisition when the incentive was nine 45-mgm. pellets for one group of rats, five such pellets for a second group, and, for a third group, nine pellets on a random half of the trials and one on the other half. Thus, the first two groups were trained according to the absolute method, the third according to the differential. Over-all running speed (60 daily trials) was greater for the first group than for either of the other two. Inspection of their Fig. 2 (p. 264) indicates similar rates of approach to terminal speeds in all cases.

In a follow-up study, Logan, Beier, and Kincaid (1956) report the extinction results of this and a related ex-

periment. For the first experiment, none of the three groups differed in their relative rates of extinction. In the related experiment, extinction followed 60 acquisition trials with nine pellets on each trial, nine vs. zero each on a random half, or nine vs. one. In this case the two varied-amount (differential method) groups were superior to the constant-amount group. The conclusion would seem to be that relative rate of extinction is retarded by previous comparison of incentive values.

Lawson (1957) assessed the effects of incentive magnitude on discrimination learning by having rats learn two problems concurrently, a black-gray discrimination and a white-gray discrimination, gray being negative in both instances. For some *Ss* the reward was the same for correct responses on both problems, either five pellets or one pellet (absolute groups). For other *Ss* the reward was five pellets on one problem and one on the other (differential groups). The two absolute groups did not differ significantly from each other in terms of errors over a constant number of trials. In contrast, the differential groups showed significantly fewer errors on the large reward discrimination than on the small reward one.

Two studies similar in design to Lawson's are reported by Schrier (1956b, 1958). In both studies, monkeys were divided into a differential and several absolute groups. In various phases of each experiment, differential *Ss* received experience with 1, 2, 4, or 8 food pellets. In contrast, members of each of four absolute groups received experience with only one of the four magnitudes. In the first study (1956b), effects were measured on the latency of response to a single object covering a baited food well. By either method, performance improved as incentive amount increased. However, the absolute and differential methods did not produce differences in over-all performance, and the slopes of the functions relating incentive magnitude to latency did not differ significantly. In the second study (1958), effects on errors during a series of discrimination problems were assessed. Again, the two groups did not differ in over-all performance. However, the slope of the incentive function was significantly greater for the differential method than for the absolute method.

Acquired Reward Value as a Function of Incentive Magnitude

Experimental studies of secondary reinforcement were recently reviewed by Myers (1958). On the basis of the evidence he discussed (D'Amato, 1955; Hopkins, 1955; Lawson, 1953), Myers concluded that, " . . . the effect of amount of reward is so slight, that it can only be detected when the *S* is forced to choose between secondary reinforcers previously associated with different sized rewards" (p. 294). The finding was that an effect appeared only when training was by the differential method, but not the absolute.

Two studies have appeared since Myers' review. In the first of these Lawson (1957) compared the absolute and differential methods within a single design, in their effects on secondary reward strength. By neither method was a differential effect of primary reward amount on secondary reward strength demonstrated.

However, a study by Butter and Thomas (1958), using the absolute method, indicates that perhaps Myers' conclusion should be revised.

During training, click was paired with the presentation of either 8% or 16% sucrose solution, and during testing, bar pressing produced the click. The 24% group emitted significantly more responses during the test than did the 8% group, indicating a differential effect of magnitude of primary reward on the strength of secondary reward. Butter and Thomas suggest that the earlier failures with the absolute method might be due to the possibility that the primary amounts were too large, toward the asymptote of the primary reward amount function.

Actually, it does not look as if any clear-cut conclusions can be reached until more research has been completed, for example using a design similar to Lawson's (1957), allowing direct comparison of the absolute and differential methods. Perhaps the safest statement that can now be made is the empirical generalization that amount of primary reward does affect strength of secondary reward, and that this effect is more likely to be evidenced if the differential method of incentive presentation is used.

Finally, this section might be concluded by mention of Wolfe's original experiments on token-reward (1936). Wolfe used the differential method in which selection of a blue token yielded two grapes, a white token yielded one grape, and a brass token yielded nothing. Three chimpanzees learned to consistently choose the blue token, demonstrating that secondary reward values can be discriminated, just as primary reward values can.

Summary and Conclusions

Perhaps the most significant finding of this review is the relatively high degree of consistency of results. The facts seem well ordered, and where contradictory results are found, such disagreements can usually be fairly readily explained in terms of procedural differences. The empirical laws are now before the theorist, for him to explain in whatever logically consistent way he can devise. Because the findings seem so well ordered, he should perhaps have less trouble than in other aspects of behavior theory.

We may conclude by bringing together the various empirical generalizations scattered throughout this paper, and offering some suggestions for further research. The following generalizations have emerged from the review of the evidence on incentive magnitude:

1. With the absolute method, quantitative variation in incentives has no apparent effect on rate of learning. A possible exception involves the manipulation of concentration of sugar solutions. A second exception is that learning is more rapid with any positive incentive amount than with zero amount.

2. Asymptotic performance is an increasing function of incentive magnitude. The function is negatively accelerated, and possibly logarithmic, although under some circumstances, nonmonotonicity appears.

3. These asymptotic differences reflect a direct effect on the level of performance, rather than an indirect effect based on differences in amount learned.

4. Magnitude of reward affects resistance to extinction indirectly through differences in terminal level of rewarded performance.

5. There exists conflicting evidence on the interaction of incentive magnitude with drive level.

6. None of the mechanisms whereby behavioral results of quantitative variation are effected are necessary.

7. Sufficient mechanisms seem to be amount of consummatory activity and stimulation from the incentive associated with consummatory activity. In the former case, it is the *rate* of consummatory activity which seems paramount. (This set of conclusions is based upon exclusion from consideration of administration of nutritive substances other than peripherally.)

8. With the differential method, incentive magnitude affects learning, but only after *S*s have learned to discriminate amounts.

9. The asymptotic incentive function is steeper when training is by the differential method than by the absolute method.

10. Acquired reward value is an increasing function of magnitude of primary reward.

To these conclusions might be added a comment on the use of time-dependent measures as measures of performance and time-independent measures as measures of learning. In the section on learning vs. performance effects by the absolute method, it was concluded that incentive magnitude affects performance but not learning. And it was typically time-dependent measures that were affected by incentive magnitude, and not time-independent measures. This correlation supports the distinction, and a more widespread adoption is recommended.

This multiplicity of conclusions does not imply that the study of incentive magnitude is a "closed" area. The conclusions are of course subject to modification, and at present some must be accepted more tentatively than others. In particular, further research seems to be most urgently needed in connection with the following problems: (*a*) further specification of the characteristics of the consummatory response as a determiner of amount of reinforcement effects; (*b*) further attention to the development of incentive magnitude discrimination (How does this type of discrimination develop? Does it follow a different course from problem-learning?); (*c*) further specification of the conditions determining contrast effects, and the relation between contrast effects obtained under the absolute and differential methods; and (*d*) the nature and conditions of the interaction of incentive magnitude, however defined, with other variables.

REFERENCES

BLODGETT, H. C. The effect of introduction of reward upon the maze performance of rats. *Univer. Calif. Publ. Psychol.*, 1929, **4**, 113–134.

BUTTER, C. M., & THOMAS, D. R. Secondary reinforcement as a function of the amount of primary reinforcement. *J. comp. physiol. Psychol.*, 1958, **51**, 346–348.

COCKRELL, J. T. Operant behavior of white rats in relation to the concentration of a non-nutritive sweet substance used as reinforcement. Unpublished doctoral dissertation, Univer. of Indiana, 1952.

COLLIER, G. Interaction of factors governing amount of reinforcement function. Paper read at Midwest. Psychol. Ass., Detroit, May, 1958.

COLLIER, G., & SISKEL, M. Performance as a joint function of amount of reinforcement and interreinforcement interval. *J. exp. Psychol.*, 1959, **57**, 115–120.

CONRAD, D. G., & SIDMAN, M. Sucrose concentration as reinforcement for lever pressing by monkeys. *Psychol. Rep.*, 1956, **2**, 381–384.

COPPOCK, H. W., & CHAMBERS, R. M. Reinforcement of position preference by automatic intravenous injections of glucose. *J. comp. physiol. Psychol.*, 1954, **47**, 355–357.

COWLES, J. T., & NISSEN, H. W. Reward expectancy in delayed responses of chimpanzees. *J. comp. Psychol.*, 1937, **24**, 345–358.

COVER, R. A. The effect of magnitude of reward and degree of deprivation on the acquisition and performance of a complex

maze habit. Unpublished doctoral dissertation, Univer. of Buffalo, 1953.

CRESPI, L. P. Quantitative variation of incentive and performance in the white rat. *Amer. J. Psychol.*, 1942, **55**, 467–517.

CRESPI, L. P. Amount of reinforcement and level of performance. *Psychol. Rev.*, 1944, **51**, 341–357.

D'AMATO, M. R. Secondary reinforcement and magnitude of primary reinforcement. *J. comp. physiol. Psychol.*, 1955, **48**, 378–380.

DAVIS, R. T. Problem-solving behavior of monkeys as a function of work variables. *J. comp. physiol. Psychol.*, 1956, **49**, 499–506.

DENNY, M. R., & KING, G. F. Differential response learning on the basis of differential size of reward. *J. genet. Psychol.*, 1955, **87**, 317–320.

DUFORT, R. H., & KIMBLE, G. A. Changes in response strength with changes in the amount of reinforcement. *J. exp. Psychol.*, 1956, **51**, 185–191.

ELLIOTT, M. H. The effect of change of reward on the maze performance of rats. *Univer. Calif. Publ. Psychol.*, 1928, **4**, 19–30.

FAY, J. C., MILLER, J. D., & HARLOW, H. F. Incentive size, food deprivation, and food preference. *J. comp. physiol. Psychol.*, 1953, **46**, 13–15.

FEHRER, ELIZABETH. Effects of amount of reinforcement and of pre- and postreinforcement delays on learning and extinction. *J. exp. Psychol.*, 1956, **52**, 167–176.

FESTINGER, L. Development of differential appetite in the rat. *J. exp. Psychol.*, 1943, **32**, 226–234.

FITTS, P. M. The effect of a large and a small reward as indicated by the resistance-to-extinction curve for the rat. *Psychol. Bull.*, 1940, **37**, 429–430. (Abstract)

FLETCHER, F. M. Effects of quantitative variation of food-incentive on the performance of physical work by chimpanzees. *Comp. Psychol. Monogr.*, 1940, **16** (82).

FURCHTGOTT, E., & RUBIN, R. D. The effect of magnitude of reward on maze learning in the white rat. *J. comp. physiol. Psychol.*, 1953, **46**, 9–12.

GAGNÉ, R. M. The effect of spacing of trials on the acquisition and extinction of a conditioned operant response. *J. exp. Psychol.*, 1941, **29**, 201–216.

GANTT, W. H. The nervous secretion of saliva: The relation of the conditioned reflex to the intensity of the unconditioned stimulus. *Amer. J. Physiol.*, 1938, **123**, 74. (Abstract)

GREENE, J. E. Magnitude of reward and acquisition of a black-white discrimination habit. *J. exp. Psychol.*, 1953, **46**, 113–119.

GRINDLEY, G. C. Experiments on the influence of the amount of reward on learning in young chickens. *Brit. J. Psychol.*, 1929, **20**, 173–180.

GUTTMAN, N. Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *J. exp. Psychol.*, 1953, **46**, 213–224.

GUTTMAN, N. Equal-reinforcement values for sucrose and glucose solutions compared with equal-sweetness values. *J. comp. physiol., Psychol.*, 1954, **47**, 358–361.

HARLOW, H. F. The formation of learning sets. *Psychol. Rev.*, 1949, **56**, 51–65.

HARLOW, H. F., & MEYER, D. R. Paired comparison scales for monkey rewards. *J. comp. physiol. Psychol.*, 1952, **45**, 73–79.

HELLYER, S. The duration of the consummatory response as a variable in amount of reinforcement studies. Unpublished doctoral dissertation, Univer. of Indiana, 1953.

HEYER, A. W. Studies in motivation and retention: IV. The influence of dehydration on acquisition and retention of the maze habit. *Comp. Psychol. Monogr.*, 1951, **20**(4), 273–286.

HOPKINS, C. O. Effectiveness of secondary reinforcing stimuli as a function of the quantity and quality of food reinforcement. *J. exp. Psychol.*, 1955, **50**, 339–342.

HUGHES, L. H. Saccharine reinforcement in a T maze. *J. comp. physiol. Psychol.*, 1957, **50**, 431–435.

HULL, C. L. *Principles of behavior*. New York: Appleton-Century-Crofts, 1943.

HULL, C. L. *A behavior system*. New Haven: Yale Univer. Press, 1952.

HULSE, S. H. Amount and percentage of reinforcement and duration of goal confinement in conditioning and extinction. *J. exp. Psychol.*, 1958, **56**, 48–57.

HUTT, P. J. Rate of bar pressing as a function of quality and quantity of food reward. *J. comp. physiol. Psychol.*, 1954, **47**, 235–239.

JENKINS, W. O., & CLAYTON, FRANCES L. Rate of responding and amount of reinforcement. *J. comp. physiol. Psychol.*, 1949, **42**, 174–181.

KAGAN, J. Differential reward value of incomplete and complete sexual behavior. *J. comp. physiol. Psychol.*, 1955, **48**, 59–64.

KLING, J. W. Speed of running as a function of goal-box behavior. *J. comp. physiol. Psychol.*, 1956, **49**, 474–476.

LAWRENCE, D. H., & MILLER, N. E. A positive relationship between reinforcement and resistance to extinction produced by removing a source of confusion from a technique

that had produced opposite results. *J. exp. Psychol.*, 1947, **37**, 494–509.

LAWSON, R. Amount of primary reward and strength of secondary reward. *J. exp. Psychol.*, 1953, **46**, 183–187.

LAWSON, R. Brightness discrimination performance and secondary reward strength as a function of primary reward amount. *J. comp. physiol. Psychol.*, 1957, **50**, 35–39.

LEARY, R. W. Homogeneous and heterogeneous reward of monkeys. *J. comp. physiol. Psychol.*, 1958, **51**, 706–710.

LEWIS, D. J., & COTTON, J. W. Learning and performance as a function of drive strength during acquisition and extinction. *J. comp. physiol. Psychol.*, 1957, **50**, 189–194.

LOGAN, F. A., BEIER, EILEEN M., & ELLIS, R. A. Effect of varied reinforcement on speed of locomotion. *J. exp. Psychol.*, 1955, **49**, 260–266.

LOGAN, F. A., BEIER, EILEEN M., & KINCAID, W. D. Extinction following partial and varied reinforcement. *J. exp. Psychol.*, 1956, **52**, 65–70.

MCKELVEY, R. K. The relationship between training methods and reward variables in brightness discrimination learning. *J. comp. physiol. Psychol.*, 1956, **49**, 485–491.

MAHER, WINIFRED B., & WICKENS, D. D. Effect of differential quantity of reward on acquisition and performance of a maze habit. *J. comp. physiol. Psychol.*, 1954, **47**, 44–46.

METZGER, R., COTTON, J. W., & LEWIS, D. J. Effect of reinforcement magnitude and order of presentation of different magnitudes on runway behavior. *J. comp. physiol. Psychol.*, 1957, **50**, 184–188.

MEYER, D. R. The effects of differential rewards on discrimination reversal learning by monkeys. *J. exp. Psychol.*, 1951, **41**, 268–274.

MEYER, D. R., & HARLOW, H. F. Effects of multiple variables on delayed response performance by monkeys. *J. genet. Psychol.*, 1952, **81**, 53–61.

MICHELS, K. M. Response latency as a function of the amount of reinforcement. *Brit. J. Anim. Behav.*, 1957, **5**, 50–52.

MILLER, N. E., & KESSEN, M. L. Reward effects of food via stomach fistula compared with those of food via mouth. *J. comp. physiol. Psychol.*, 1952, **45**, 555–564.

MYERS, J. L. Secondary reinforcement: A review of recent experimentation. *Psychol. Bull.*, 1958, **55**, 284–301.

NISSEN, H. W., & ELDER, J. H. The influence of amount of incentive on delayed response performance of chimpanzees. *J. genet. Psychol.*, 1935, **47**, 49–72.

PEREBOOM, A. C. An analysis and revision of Hull's theorem 30. *J. exp. Psychol.*, 1957, **53**, 234–238. (a)

PEREBOOM, A. C. A note on the Crespi effect. *Psychol. Rev.*, 1957, **64**, 263–264. (b)

PEREBOOM, A. C., & CRAWFORD, B. M. Instrumental and competing behavior as a function of trials and reward magnitude. *J. exp. Psychol.*, 1958, **56**, 82–85.

POWELL, D. R., & PERKINS, C. C. Strength of secondary reinforcement as a determiner of the effects of duration of goal response on learning. *J. exp. Psychol.*, 1957, **53**, 106–112.

REYNOLDS, B. The acquisition of a black-white discrimination habit under two levels of reinforcement. *J. exp. Psychol.*, 1949, **39**, 760–769.

REYNOLDS, B. Acquisition of a simple spatial discrimination as a function of the amount of reinforcement. *J. exp. Psychol.*, 1950, **40**, 152–160. (a)

REYNOLDS, B. Resistance to extinction as a function of the amount of reinforcement present during acquisition. *J. exp. Psychol.*, 1950, **40**, 46–52. (b)

REYNOLDS, B., MARX, M. H., & HENDERSON, R. L. Resistance to extinction as a function of drive-reward interaction. *J. comp. physiol. Psychol.*, 1952, **45**, 36–42.

REYNOLDS, W. F., & PAVLIK, W. B. Running speed as a function of deprivation period and reward magnitude. Paper read at Midwest. Psychol. Ass., Detroit, May, 1958.

SCHRIER, A. M. Amount of incentive and performance on a black-white discrimination problem. *J. comp. physiol. Psychol.*, 1956, **49**, 123–125. (a)

SCHRIER, A. M. Effect of method of presenting varied amounts of food incentive on performance by monkeys. Unpublished doctoral dissertation, Univer. of Wisconsin, 1956. (b)

SCHRIER, A. M. Comparison of two methods of investigating the effect of amount of reward on performance. *J. comp. physiol. Psychol.*, 1958, **51**, 725–731.

SCHRIER, A. M., & HARLOW, H. F. Effect of amount of incentive on discrimination learning by monkeys. *J. comp. physiol. Psychol.*, 1956, **49**, 117–122.

SEWARD, J. P., SHEA, R. A., & ELKIND, D. Evidence for the interaction of drive and reward. *Amer. J. Psychol.*, 1958, **71**, 404–407.

SHEFFIELD, F. D., & ROBY, T. B. Reward value of a non-nutritive sweet taste. *J. comp. physiol. Psychol.*, 1950, **43**, 471–481.

SHEFFIELD, F. D., ROBY, T. B., & CAMPBELL, B. A. Drive reduction versus consumma-

tory behavior as determinants of reinforcement. *J. comp. physiol. Psychol.*, 1954, **47**, 349–354.

SHEFFIELD, F. D., WULFF, J. J., & BACKER, R. Reward value of copulation without sex drive reduction. *J. comp. physiol. Psychol.*, 1951, **44**, 3–8.

SMITH, M., & DUFFY, M. Evidence for a dual reinforcing effect of sugar. *J. comp. physiol. Psychol.*, 1957, **50**, 242–247.

SMITH, M., & KINNEY, G. C. Sugar as a reward for hungry and nonhungry rats. *J. exp. Psychol.*, 1956, **51**, 348–352.

SPENCE, K. W. *Behavior theory and conditioning*. New Haven: Yale Univer. Press, 1956.

THISTLETHWAITE, D. A critical review of latent learning and related experiments. *Psychol. Bull.*, 1951, **48**, 97–129.

TINKLEPAUGH, O. L. An experimental study of representative factors in monkeys. *J. comp. Psychol.*, 1928, **8**, 197–236.

TOLMAN, E. C., & HONZIG, C. H. Introduction and removal of reward, and maze performance in rats. *Univer. Calif. Publ. Psychol.*, 1930, **4**, 257–275.

VERHAVE, T. The effects of varying the concentration of a sucrose solution used as a reinforcing agent in a chaining situation. Unpublished doctoral dissertation, Columbia Univer., 1955.

WIKE, E. L., & BARRIENTOS, G. Selective learning as a function of differential consummatory activity. *Psychol. Rep.*, 1957, **3**, 255–258.

WOLFE, J. B. Effectiveness of token-rewards for chimpanzees. *Comp. Psychol. Monogr.*, 1936, **12**, No. 60.

WOLFE, J. B., & KAPLON, M. D. Effect of amount of reward and consummative activity on learning in chickens. *J. comp. Psychol.*, 1941, **31**, 353–361.

YOUNG, P. T., & SHUFORD, E. H. Intensity, duration, and repetition of hedonic processes as related to acquisition of motives. *J. comp. physiol. Psychol.*, 1954, **47**, 298–305.

YOUNG, P. T., & SHUFORD, E. H. Quantitative control of motivation through sucrose solutions of different concentrations. *J. comp. physiol. Psychol.*, 1955, **48**, 114–118.

ZEAMAN, D. Response latency as a function of the amount of reinforcement. *J. exp. Psychol.*, 1949, **39**, 466–483.

(Received May 1, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 2, 1960

THE PARAMORPHIC REPRESENTATION OF CLINICAL JUDGMENT[1]

PAUL J. HOFFMAN
University of Oregon

The primary task of clinical diagnosis is that of collecting, evaluating, and assimilating information with respect to the patient. The starting point is the information itself; this may be in the form of laboratory test results, biographical data, scores on psychological tests, manifest symptoms, or other observables. The end result is a judgment; this may take the form of a recommendation concerning treatment or discharge, a decision that certain other data are necessary before final judgment is made, or a classification of the patient into a diagnostic category. What intervenes between beginning and end is, for each clinician, a quite complex idiosyncratic process. It is the purpose of this paper to demonstrate that the process is capable of rigorous investigation and description.

THE MENTAL PROCESS

In dealing with the manner in which clinicians utilize information at their disposal to arrive at judgments or decisions, it may appear that investigations would be concerned primarily with mental processes; and since mental processes have often been equated to or placed within the realm of subjective experience, it would be well to make one or two observations for purposes of clarification. In the first place, the term *mental process* is often directly equated with subjective experience. But as private experience, the mental process is not observable. Hence, acceptance of this definition places the process beyond the realm of legitimate scientific inquiry, except as it may be *inferred* from observable phenomena such as verbal responses And since no criterion can exist for the validation of inferences concerning subjective experience, the inferences are simply ways of finding agreement in the use of language or other symbolic responses between the subject and the observer. If an observer makes "good" inferences concerning a subject, this means at most that a consensus exists between them with respect to the symbolic behavior involved. Understanding of the mental process qua subjective experience can never go beyond this level.

On the other hand, mental process may be alternately defined. It may be considered as a physical (e.g., neurological, biochemical) event capable of direct observation, i.e., using electrophysiological, neurophysiological, and similar techniques. To be sure, these techniques have so far yielded little that satisfactorily describes cognitive mental functioning, and this is perhaps unfortunate. It does not follow that the approach is sterile. Improvement in the techniques of measurement and in the application of more explanatory models may ultimately result in great

[1] This investigation was supported by a Public Health Service research grant (M2097-C1) from the National Institute of Mental Health.

The analysis of much of the data referred to in this report was made possible through the facilities of the Division of Counseling & Testing Services, University of Washington, and the Western Data Processing Center at the University of California, Los Angeles.

progress, even though this may seem remote at the present time. This second definition is not unreasonable by any known standard, and it may surely encourage productive research and a resulting clarification of basic issues. But it is perhaps too early to say.

This brings us to the third sense in which the term mental process may be employed. It should first be pointed out that any realm of scientific investigation is designed to provide, among other things, a useful level of objective description. Direct observation, testing, instrumentation, and other related techniques are steps in this direction. When properly employed within a theoretical framework they seek to describe relationships between events or phenomena. The problem of describing judgment can similarly be considered to be one which interposes a set of techniques and a theoretical system between two sets of observables. Thus it is possible to "describe" the kinds of mental activity usually characterized as cognitive by means of mathematical models. One may thereby approach a level of description which is at least equal to that of other competitors in some respects, and certainly superior in other ways. That is to say, in controlled situations wherein the input (information) and the output (judgment) are known or capable of quantification, one may postulate functional relationships between input and output and assess their adequacy by determining the accuracy with which each is capable of predicting judgment. The present paper is directed at this level of description. The term mental process refers simply to a functional relationship which accounts for consistencies in response to divergent stimulus (information) patterns. It is thus a set of intervening variables, nothing more.

A question which immediately arises from the foregoing discussion is that of the adequacy with which it is possible to describe the mental processes underlying clinical judgment. In answer it may be said that the process is adequately described when a particular mathematical model quite effectively predicts judgments for any given set of information. This is consistent with the scientific meaning of the word "description," although considerations such as simplicity, generality, and the testability of derivations must be kept in mind. A major problem in the understanding of judgment will in this paper be considered to be that of formulating the kind of model which is sufficiently predictive, yet useful as a vehicle for approaching other related problems in the area of judgment. Different kinds of models need to be discussed, compared, and evaluated, and empirical findings are of course necessary. Subsequent sections of this paper will consider specific models for judgment. One, the linear model, is relatively simple; another to be described is somewhat more complex. Following this, some empirical findings will be offered as illustrative of the research opportunities which unfold.

Before beginning the discussion of specific models, however, it becomes necessary to justify certain restrictions that must be imposed in order that meaningful inquiry may be made into the judgment process. The restrictions center in the nature of the information available to the judge or clinician and upon which the judgment is contingent. The restrictions do not seriously impair the realism of the judgment situation as long as one

can bring some ingenuity to bear upon certain problems of quantification, but even this point of view may be objectionable to some. For completeness, therefore, and in order to provide early insight into the experimental procedures used in the empirical studies to be later described, attention may now be directed to the problem of the nature of the information available to the judge.

The Information

The information upon which clinical assessment is based may commonly be expected to include anything or everything, depending upon the training and inclination of the diagnostician, and depending upon what is available or easily obtainable. The lack of control over such information may be considered an asset or a liability, depending upon one's orientation, and the accuracy of judgment may or may not be enhanced through the inclusion of nonquantitative data (Ullmann & Berkman, 1959; Holt, 1958; Luft, 1950; Meehl, 1954) depending upon the judgment domain or situation.

What seems certain, regardless of the outcomes of empirical research, is that the uncontrolled use of clinical data, whether or not it exists in quantitative form, makes clinical assessment an artistic venture. It must of course remain a matter of one's values as to whether this is wise. What seems equally certain is that any seriously conducted scientific study of judgment which has as its purpose the description of the method of combination used by the judge must take place in controlled settings, i.e., in such a way that the amount, kind, and nature of the information available to the clinician or judge can be completely specified in objective terms.

Controlling the judgment task to this degree has its advantages and its liabilities. On the one hand, restricting the situation as described assures that each person is evaluated with respect to the same information. Ambiguous and equivocal cues are removed, and all judges are thereby certain to have at their disposal the same information and no more. The inferences made beyond this point are thus certain to have their origins in the data provided. The major problem, that of describing the idiosyncratic method of combination and weighting of this information by the clinician, is thereby clearly defined. Clinical judgments are, of course, often made in settings wherein the kinds of information available may include interview and projective test impressions, etc. In addition, the kind of information available may vary considerably from one patient to the next. This may be said to pose a limitation to the situation described. Such information may be important in judgments, a point perhaps best made by Holt (1958), but unstructured clinical judgment situations nonetheless make the contribution of such information experimentally impossible to assess. The problem here under investigation would, as a result, cease to be a scientific problem altogether.

Let us therefore consider that the situation in which a clinician makes evaluations of patients is restricted in the following ways: (*a*) the information available is reduced to a set of variables with respect to which all patients in the sample are evaluated; (*b*) the information is expressed in numbers or in categorical responses; and (*c*) each variable satisfies as a minimum the properties of an ordinal scale. This scheme leads to a set of numbers or classifi-

cations for each patient, where each number or class represents the degree to which a characteristic, trait, symptom, or biographical factor is present. One example of the situation satisfying these restrictions is that of a symptom check list; another is that of a test profile. Rating scale data are likewise permissible, as would be combinations of these types.

Having objectified the data upon which the judgments are to be based, we may now turn to a consideration of the model.

The Linear Model

The linear model is one in which judgments are described as a simple weighted sum of the values of the information available. For a given clinician judging a number of people, we let J represent the judgment and consider it as a dependent variable. The dimensions of information are designated by Xs. These will, of course, be independent variables. If there are k sources of information, the linear additive model can be described as follows:

$$J = f(X_i)$$
$$i = 1, 2, \cdots k$$

Since we are interested in a weighted sum of the X_i we may write

$$J = A_0 + A_1X_1 + A_2X_2 + \cdots + A_kX_k.$$

If the A_i are so chosen as to yield the best possible weighted sum, i.e., so that the composite scores correlate maximally with J, the model is equivalent to a linear multiple regression equation wherein the weights to be applied to the independent variables are so chosen as to minimize the error in estimating an actual dependent variable from the weighted composite.

Application of multiple regression procedures to the problems of judgment has been suggested by Brunswik (1947), and by Hammond (1955). Todd (1954) reports a study using regression coefficients and the multiple correlation coefficient for a description of the clinical judgment process, where the task was to judge intelligence from a selected number of Rorschach signs. While such studies provide interesting implications, it should be stressed that there are serious limitations with respect to the interpretation of results; limitations which may be minimized or overcome only through a detailed examination of the rationale underlying the model, and through reformulations or revisions of the model, should this be necessary. So as to insure the appropriateness of the linear model as a device for characterizing the judgment process, we consider in detail some of its properties, and provide the particular reformulations where necessary.

In the first place, and by virtue of the experimental control employed in the collection of the data, the only source of reliable judgment variance is from the information supplied. This is in objective form, e.g., it appears as a number, a designated category, a position along a continuum, etc. Often these data appear as test scores on a set of protocols being judged. Assuming that a judge combined the information in linear additive fashion, the multiple regression analysis will be quite effective as a tool for describing the judgment process; i.e., the set of regression weights when applied to the corresponding predictors can quite properly serve as a model for judgment. Thus, the adequacy of the linear model can be assessed by inspection of the magnitude of multiple R. If the judge integrates data in additive fashion as opposed to

configurational or pattern analysis, the linear multiple correlation will approach unity when corrected for attenuation. Lesser values of R suggest progressively lesser utility for the linear model.

Secondly, it may be noted that the regression weights signify, with certain limitations, the emphasis or importance attached to each of the predictor variables by the judge. Large coefficients mean, empirically, that the corresponding predictors can account for large proportions of the variance of judgment; and a predictor with a small beta coefficent contributes little beyond the contribution of other predictors. In practice, characterization of the judgment process by means of beta coefficients has three limitations: (*a*) since *J*s differ with respect to the size of their multiple R, direct comparisons of sets of beta coefficients between *J*s is not meaningful; (*b*) beta coefficients do not account for all the predictable variance; and (*c*) beta coefficients do not allow for the assessment of the *independent* contribution of each predictor. What would be more appropriate would be a set of weights which are comparable from one *J* to the next, which are capable theoretically of accounting for all of the predictable variance, and which carry exact interpretation in terms of components of variance.

Relative Weights

The formulation that is required is fortunately not difficult. Beta weights (β_{oi}) can be converted into a set of relative weights (w_{oi}) which have all the advantages described. We show first that the variance of predicted scores (which in this paper refers to predicted judgments) can be partitioned into two sources; one a sum of squared beta coefficients, and the other a residual of weighted covariances.

Let

x'_o = the predicted score for an individual (protocol) in reduced standard form.

x_i = the standard score of the ith predictor (on the protocol).

β_i = the beta coefficient for the ith predictor.

$$x'_o = \beta_{o1}x_1 + \beta_{o2}x_2 + \cdots + \beta_{oi}x_i + \cdots + \beta_{ok}x_k$$

or

$$x'_o = \sum_{i=1}^{k} \beta_{oi}x_i$$

$$\sigma^2_{x'_o} = \frac{\sum x'^2_o}{N} = \frac{1}{N}\sum^{N}\left[\sum_{i=1}^{k}\beta_{oi}x_i\right]^2$$

The term in parenthesis is a weighted variance-covariance matrix. It can be divided as follows:

$$\left[\sum_{i=1}^{k}\beta_{oi}x_{oi}\right]^2 = \sum_{i=1}^{k}\beta^2_{oi}x_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{k-1}\beta_{oi}\beta_{oj}x_ix_j (i>j)$$

The first quantity on the right, when squared and averaged over individuals, yields the squares of the β coefficients. Thus:

$$\frac{\sum^{N}\sum_{i=1}^{k}\beta^2_{oi}x_i^2}{N} = \sum_{i=1}^{k}\beta^2_{oi}\sigma^2_{x_i} = \sum_{i=1}^{k}\beta^2_{oi}$$

since the x_i are standard scores.

Similarly, the second quantity is a weighted sum of the intercorrelations among the predictors. Thus:

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{k}\sum_{j=1}^{k-1}\beta_{oi}\beta_{oj}x_i x_j}{N} = \sum_{i=1}^{k}\sum_{j=1}^{k-1}\beta_{oi}\beta_{oj}r_{ij} \qquad (i>j)$$

Therefore,

$$\sigma^2_{o'_o} = \sum_{i=1}^{k}\beta^2_{oi} + \sum_{i=1}^{k}\sum_{j=1}^{k-1}\beta_{oi}\beta_{oj}r_{ij}$$

It follows that the variance of predicted scores is described by a simple sum of squared beta coefficients if and only if the covariance terms vanish. One special case in which this will be true is that of orthogonal predictors.

Relative weight, w_{oi}, is defined as follows: First we note that

$$\sqrt{\beta_{01}r_{01}+\beta_{02}r_{02}+\cdots+\beta_{ok}r_{ok}} = R_{0\cdot 12\cdots k}$$

or

$$\sqrt{\sum_{i=1}^{k}\beta_{oi}r_{oi}} = R_{0\cdot 12\cdots k}$$

Squaring both sides and dividing by R^2, we get

$$\sum_{i=1}^{k}\frac{\beta_{oi}r_{oi}}{R^2_{0\cdot 12\cdots k}} = 1$$

Therefore, in interpreting by independent components of variance, we express *relative weight* as

$$w_{oi} = \frac{\beta_{oi}r_{oi}}{R^2_{0\cdot 12\cdots k}}$$

where

β_{oi} = the beta coefficient for the ith predictor

r_{oi} = the validity coefficient (correlation with judgment) of the ith predictor

$R^2_{0\cdot 1,2\cdots k}$ = the squared multiple correlation coefficient reflecting the best linear combination of the k predictors in prediction of judgments.

Finally, a description of the judgment process by means of linear regression procedures and relative weights allows one to go on to studies of varied sorts. Judges may be compared and contrasted with respect to their characteristic equations; and differences among judges may be related to training, personality, and other factors that could conceivably affect the utilization of data. Many other problems immediately suggest themselves.

The linear model may effectively be able to predict (or describe) clinical judgments to a very considerable degree, but there may be other situations for which linear models are not appropriate—just as there must be many judges for whom more complex models are necessary. Let us now turn our attention to a second type of model.

Configurational Models

In very general terms, the configurational model can be described as

$$J = f(X_1, X_2, X_3, \cdots X_k)$$

wherein the exact functional relationship involving the k independent variables may be described in any of a number of ways. As an example, let us consider a particular type of function, one which shall be referred to as an *interaction model*. The interaction model describes judgment as an appropriately weighted composite of all possible first order interactions of the predictors. Thus we may write

$$J = A_o + \sum_{i=1}^{k}\sum_{j=1}^{k-1}A_{ij}X_iX_j \qquad (i>j)$$

The inclusion of interaction terms in a model takes account of the possibility that for a particular judge the interpretation of one item of information may be contingent upon a second. By extension, other interaction models suggest themselves. Thus it is possible to include terms involving higher order interactions or other postulated functional relationships among the information variables. Such models represent configural judgments in a quite proper sense. As the postulated interrelationships become more complex, the judgment becomes less dependent upon any single category of information and less dependent upon a simple weighted sum. Instead, the judgment comes to depend upon the configural properties of the profile, and these may approach a high degree of uniqueness. Correspondingly, the person judged is increasingly evaluated not so much with respect to a reference group of others earning the same score, but rather with respect to the *pattern* of his scores. And parenthetically, as in the case of the additive model, some of these may be judged as functionally equivalent though the scale scores differ markedly.

Other configurational models may be postulated, some of which do not involve interaction among the independent variables, but instead require a transformation of one or more of them into a different set of units. The apparent need for such transformation arises from the assumption that clinicians seldom believe that linear functions best describe relationships between information and characteristics being judged. It may, in fact, be closer to the truth that, at least for some classes of information, extreme scores are more decisive in judgment than are scores in the middle range. And clinicians will of course be expected to differ amongst themselves. For some, scores above or below some arbitrary value may carry no added significance whatsoever. In selecting configurational models for study, it therefore becomes necessary to take into account the great individual differences which may exist among clinicians and to construct that type of model which appears most promising. The application of such a model to a well-controlled situation may ultimately be capable of accounting for all but a trivial fraction of the variance of judgments.

Suppressor Effects

One of the difficulties inherent in the interpretation of beta coefficients becomes apparent when it is desired to make some statement concerning either causality or relative contribution. A beta may be high because the predictor with which it is associated correlates highly with the criterion and is relatively independent of other predictors. But it may also be high, even though the variable with which it is associated is itself a very poor predictor of the criterion, so long as its correlation with other valid predictors is sufficiently high. The point has been made in most statistics textbooks and is discussed in a few recent journal publications (e.g., Lubin 1957). The example presented in McNemar (1955) will serve as illustrative. Assume the following:

$$r_{01} = .400$$

$$r_{02} = .000$$

$$r_{12} = .707$$

where x_0 is a judgment criterion, x_1 and x_2 are predictors (information).

Solution of this matrix leads to $\beta_{01} = .800$, $\beta_{02} = -.566$, $R_{0.12} = .566$.

Quite evidently, the second predictor is a suppressor. It carries negative weight because it accounts for variance in the first predictor that is independent of the criterion. Beta coefficients would not adequately describe the judgment process for this case, since β_{02} would be $-.566$ even if this second predictor were unavailable as information to the judge! Indeed, it is possible, after having administered a set of protocols to a judge to add additional predictors to the correlation matrix in any arbitrary manner, and some of these might well yield significant betas.

The use of relative weights obviates this difficulty. In the example described,

$$w_{01} = \frac{(.400)(.800)}{(.566)^2} = 1.000$$

$$w_{02} = \frac{(.000)(-.566)}{(.566)^2} = .000$$

and it therefore becomes clear that a predictor must itself correlate significantly with the judgment (criterion) in order to obtain a significant relative weight.

A second point may be raised in this connection, and with reference generally to the interpretation of relative weights. Is it possible for a predictor to acquire a significant relative weight simply by virtue of its being correlated with a valid predictor? If, for example, one is asked to judge intelligence from high school rating and father's IQ (these two predictors being highly correlated), might not father's IQ turn out to have a significant relative weight by virtue of this correlation, even though the judge ignored it on the protocols?

This problem can be seen most easily in the context of partial correlation, and with reference to the extreme case. If the validity coefficient r_{01} is attributable exclusively to a second predictor,

$$r_{01.2} = 0,$$

and since

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1-r^2_{02}}\ \sqrt{1-r^2_{12}}}$$

then

$$r_{01} - r_{02}r_{12} = 0$$

Further:

$$\beta_{01.2} = \frac{r_{01} - r_{02}r_{12}}{1-r^2_{12}}$$

and:

$$\beta_{01.2} = 0$$

It would appear therefore that relative weights as defined do in fact provide meaningful descriptions of the process of judgment, i.e., with respect to the relative importance of the various items of information available to the judge, and without the kinds of spurious effects often associated with multiple regression procedures. Further, it would seem that a general mathematical proof for this conclusion is not difficult to develop. These considerations apply both to linear and configurational models.

Equivalence Among Models

A special problem becomes apparent when it is recognized that two or more models may be capable of accounting for judgment variance with equal efficiency. Consider, for example, a given model which is highly accurate in predicting judgments from the information given. In this sense we may be said to have characterized or "described" the judgmental process, but one im-

portant qualification is necessary. Even in the hypothetical situation in which prediction is perfect, one cannot conclude that the mental process has been "discovered." By definition, of course, this point should be obvious, but it is well to point out that even among sets of mathematical relationships (models) which are ostensibly different, there may be some which are in fact equivalent with respect to explanatory power.

An example may serve to clarify this point. Let us assume that for a given judge, and for two information variables, X and Y, the judgments can be independently predicted from X and Y with 95% accuracy by the following equation:

$$J' = +\sqrt{X^2+Y^2+2XY}$$

We note that the right hand term is simply the square root of the binomial $(X+Y)^2$. Since $X+Y = +\sqrt{X^2+Y^2+2XY}$, it follows that the equation $J' = X+Y$ will account for the judgments equally as well as the expression

$$J' = +\sqrt{X^2+Y^2+2XY}$$

It is therefore no more reasonable to conclude that the judge is in fact "using" one particular combination of the information than it is to conclude that he is using the other. One would have to establish different criteria before a choice between two such representations may be intelligently made.

But this should not be a troublesome point. Mathematical models are designed to provide a scheme whereby one set of events may be satisfactorily predicted from another, and whereby testable derivations may lead to more complete theoretical understanding of the phenomena. Such models therefore constitute a level of description and explanation which suffices for scientific purposes. It is not required of models that they bear any semblance of some "actual" state of affairs, either within the organism or elsewhere, nor would this necessarily lead to a better understanding of nature.

This may be more clearly seen in relation to an example from the physical sciences. A chemist has the task of describing a substance. He performs a number of operations (or tests) on the substance and determines its chemical composition. It turns out to be a relatively simple task. The substance is described, chemically, as $CaCO_3$. But the work is not complete, for a different set of operations (or tests) might have produced a different level of description; one which may not be necessary for the chemist, but which is more suitable for other contexts. The mineralogist, for example, would perform a series of tests to measure the optical properties of the crystal, or he might examine its hardness, solubility, and other characteristics. Two crystals identified by the chemist as $CaCO_3$ might in fact be somewhat different from one another to the mineralogist, one being aragonite, and the other calcite. These two crystals do in fact have the same chemical structure, but they differ in molecular structure; this being revealed by optical and other tests. It is apparent that different levels of description are possible with regard to substances, and that each has its peculiar advantages and shortcomings.

In mineralogy, calcite is commonly described as a *paramorph* of aragonite. This word is used generally to describe a substance having crystalline structural properties which dif-

fer from those of another substance with the identical chemical composition.[2]

We have borrowed the term *paramorphic* from mineralogy and employ it in relation to representations of human judgment. The analogy may not be complete, but its limitations are not serious. The mathematical representation of the judgment process is a level of description that approaches the chemical description of minerals. The formula helps to account for or "explain" what is observed concerning certain properties or characteristics of the judge, just as the chemical formula "explains" many, though not all, properties or characteristics of the substance. In addition, the formulae are useful in making predictions concerning the outcomes of certain other tests which may later be employed. But as with chemical analysis, the mathematical description of judgment is inevitably incomplete, for there are other properties of judgment still undescribed, and it is not known how completely or how accurately the underlying process has been represented. The term "paramorphic representation," used in relation to judgment, would seem adequately to indicate this state of affairs.

The Linear Model: Representative Results

As illustrative of the significance of the methodology described above, four persons who participated as judges in studies at Oregon have been selected. Two of these made judgments of "intelligence" of 100 persons on the basis of a set of nine predictors or sources of information. The remaining two made judgments of "sociability" of 150 persons on the basis of profiles containing scores on eight selected Edwards Personal Preference Schedule (EPPS) variables. In all cases, the judges returned after an intervening period of several days and made a second set of judgments on the same profiles. Sample profiles are shown in Fig. 1 and Fig. 2.

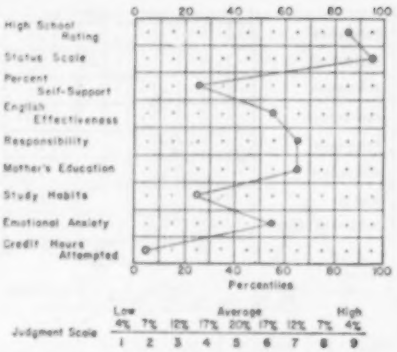

Fig. 1. A sample profile for the judgment of intelligence.

Application of standard multiple regression procedures reveals that a best linear combination of the predictor scores correlates .948 and .829 with the judgments of Judges 15 and 18, respectively.[3] But to what extent can it be said that the linear model adequately characterizes the judgment process for these judges? The reliability of judgment for these two is .876 and .836 respectively. Correction for attenuation results in

[2] More exactly, the crystal is called a paramorph when it is shown to be an alteration in crystal structure. In the example cited, aragonite may undergo change over a period of time, finally becoming calcite. It is the calcite that results from alteration of aragonite which is paramorphic.

[3] Corrected for shrinkage. Multiple *R*s computed from a cross-validation sample of protocols were .937 and .837 for these *J*s respectively.

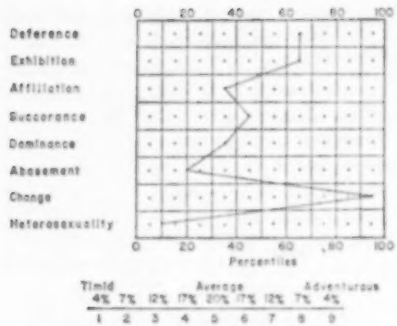

Fig. 2. A sample profile for the judgment of sociability.

coefficients which are 1.00+ and .907. Thus it may be seen, in the case of the first judge, that a linear additive combination of predictors (once a "best" set of regression weights is known) allows the prediction of judgment with virtually complete certainty. The residual or error variance is trivial in comparison with that which is predictable from the model. For the second judge the case is somewhat different. When unreliability of judgment is taken into account, there still remains 17.7% of variance which is unaccounted for by the model. It follows, therefore, that the linear model, while quite sufficient for the first judge, is less appropriate for the second.

A second type of question may be asked with respect to these judges: Is a judge able to describe with any ascertainable degree of validity the manner in which he utilizes information in arriving at his judgment? In its present form this question is probably unanswerable, but it is quite proper to ask a related one. Concerning the method of utilization of information, what correspondence exists between the verbal description offered by the judge and the description achieved by the multiple regression model?

There are some difficulties in attempting to ascertain, from the statements of the judge, a subjective impression of a cognitive process. In instances in which persons are asked to make judgments of intelligence and sociability from the sets of information that have just previously been described, it is rarely true that the judge has a high degree of confidence in statements he may make concerning the relative importance of the predictor variables. In some instances difficulties of communication emerge; when, for example, a judge finds it necessary to relate some rather complex or configurational analysis that he feels best describes his own "method of combination." One alternative is to ask the judge to distribute 100 points among the sources of information available and in such a way that this distribution reflects, to the best of his knowledge, the relative importance of those variables.

Such a task is easily understood by the judge. The method has the additional advantage of insuring that his subjective impressions are personally translated into numerical form without interpretation by a second person. Presumably, some information is lost in the process, as would be the case wherein a judge does not consider his method of combination to be adequately describable by a weighted sum of the information. Notwithstanding such attitudes, however, subjects used as judges in studies presently underway commonly report satisfaction and confidence in the procedure, and without apparent relationship to their attitudes concerning the complexity of the method of combination they believe themselves to be using.

The number of points assigned in

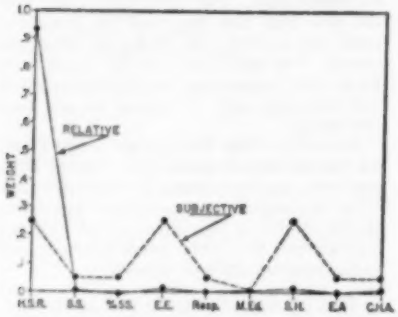

Fig. 3. Comparison of relative and subjective weights in the judgment of intelligence (Judge 15).

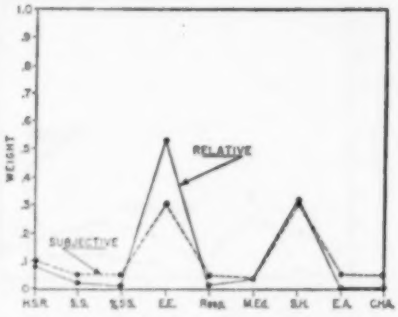

FIG. 4. Comparison of relative and subjective weights in the judgment of intelligence (Judge 18).

this way to each of the information variables will be referred to as the *subjective weight* (s_{ei}). Comparisons of subjective and relative weights are shown for the two judges of intelligence in Figs. 3 and 4. With respect to Judge 18, there is a high degree of agreement of relative and subjective weights. For Judge 15, however, there are greater discrepancies, disagreement being most pronounced in Variables 1, 4, and 7. These judges differ in the extent to which they are capable of assigning a set of numbers to the sources of information used in judgment so as to approximate the relative weights as determined by the linear model. Results from a relatively large sample of judges will be reported in a forthcoming article.

The two examples from the sociability experiment can be used to illustrate the same phenomena. The relevant information is described in Table 1 and in Figs. 5 and 6.

A Configurational Model: Representative Results

Data for one type of configurational model comes from a study by Martin (1957). On the basis of lengthy and rather complete interviews, Martin was able to obtain fairly clear statements from a set of five counselling psychologists. These statements expressed the manner in which the psychologists believed they were utilizing the set of eight EPPS variables in the prediction of sociability. Of the five judges, Psychologist D has been selected for

TABLE 1

Goodness of Fit of Linear Model for Judgments of Intelligence & Sociability

		r_{11}	R_e	Attenuated R_e	Nonlinearity[a]
Intelligence	Judge 15	.876	.948	1.00*	—
	Judge 18	.836	.829	.907	17.7
Sociability	Judge 3	.830	.901	.989	2.2
	Judge 6	.830	.770	.845	28.6

[a] % of variance unpredictable from the linear model.

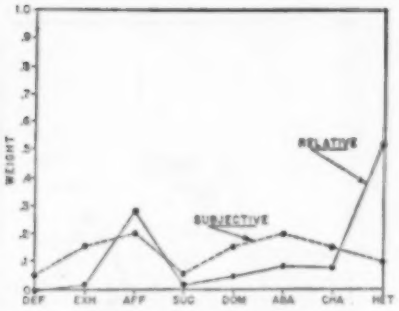

FIG. 5. Comparison of relative and subjective weights in the judgment of sociability (Judge 3).

illustrative purposes. His verbalizations with respect to the judgment of sociability follow:

As might be expected I generally look for some over-all patterning of the test variables. Although I do hold in mind certain standards and/or tendencies. In the following discussion when I refer to high I mean at least 1 *SD* above the mean. For a rating of high Sociability, i.e., 7, 8, or 9, I would generally expect at least two or three scales (Exh.+Dom.+Het.) to meet the criterion of 1*SD* above the mean with others at least in the middle range or pointing in the direction of high. I would also expect for this rating that Aba. be in the average range or in the direction of low. The more that Exh.+Dom.+Het. approach the high extreme and Aba. the low extreme the more apt I am to make a higher rating. In this concept the other four scales act in pairs. For example, if the above conditions are met and Def. and Suc. are not in either extreme the rating remains unaffected. If both are extremely low they tend to add to the rating and if both are extremely high they tend to detract from the rating of high Sociability.

The scales Chg. and Aff. are also considered as a pair but add or detract nothing to the ratings unless both are quite high or quite low. If both should be extremely low and the other conditions for a high rating are met I would suspect the reliability of the test.

In the case of a low rating (1, 2, 3) I would generally expect a somewhat opposite patterning. For example, here I would expect Aba. to be quite high with pair Suc. and Def. in the middle range or pointing in the direction of high. Similarly, I would expect at least two of the three scores (Dom., Exh., and Het.) to be in the middle or low range with none of them extremely high. The higher Aba. along with the pair Def. and Suc. and the lower the scales Dom., Het., and Exh., the lower the rating. The pair Chg. and Aff. again have little effect unless they are significantly low and then they tend to support or add to a low rating.

In rating within the average range I look for the significant scales Exh., Dom., Het., and Aba., not be extreme in either direction. While the variables are considered in relation to one another (a high score on Exh. offsets a high score on Aba.) they contribute to my final rating whether singly or in pairs.

From the foregoing description, it may be argued that the set of variables underlying Ds judgments must take into account the following considerations:

1. *Interaction.* Certain of Ds statements imply an interactive or multiplicative relationship between two predictors and the judgment criterion. A good example of this is the statement, "The scales Chg. and Aff. are also considered as a pair but add or detract nothing to the rating unless both are quite high or quite low."
2. *Nonlinearity.* Since D has stated that variables are of most importance when their values are high, and that values between $\pm 1\sigma$ of the mean are usually ignored, an exponential function should be more predictive of judgment than a linear one.

As a result, the following predictors are defined:

X_1 Abasement
X_2 Exhibitionism
X_3 Heterosexuality

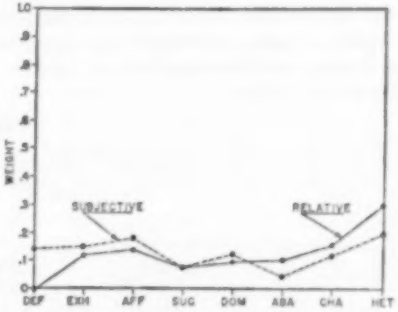

FIG. 6. Comparison of relative and subjective weights in the judgment of sociability (Judge 6).

X_4 Dominance
X_5 Change × Affiliation
X_6 Succorance × Deference

Each of the six predictors can then be subjected to a nonlinear transformation, the result of which is to correct for the stated tendency of the judge to discount predictor scores near the mean of the distribution, and to emphasize them increasingly as they become extreme, up to a limit. The resulting variables are then weighted according to least squares procedures.

As in the case of the linear model, it is possible to compute relative weights and to compare these with the weights assigned subjectively. This comparison is shown in Fig. 7. We shall discuss the data shortly.

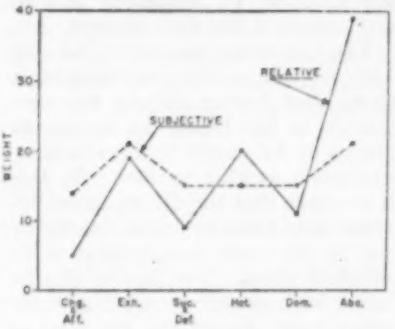

FIG. 7. Comparison of relative and subjective weights, using a nonlinear model. Clinician D (From Martin, 1957).

What classes of problems can be attacked through the technique of the configurational model? As in the case of the linear model, it furnishes a description of the relative importance (to the judge) of the various sources of information available. But greater latitude is possible. Configurational models are capable of handling the complexities and patterns believed by many to be an essential (if not "natural") part of the judgment process. Thus, one may ask whether a graduate course in psychodiagnostics or in personality assessment was effective in producing students who are "configurational" in their interpretation of case material.

A second class of problem is that of individual differences. Do persons differ in the type of model which most appropriately accounts for their judgments? If so, in what respects? And what proportion of these differences can be attributed to specific training? To personality? To intellectual characteristics?

There is a third class of problem, one which is at least of equal significance to the others mentioned. This concerns the stability of the judgment process. Can a "linear" judge be taught to be more complex? Can any judge be taught to make more efficient and more accurate use of information? Will this generalize or transfer to other judgment situations? What are the personality characteristics which differentiate the flexible judge from the unchangeable one? Some of these problems are currently under investigation and results will appear in forthcoming publications.

Finally, it may already have become apparent that the issue of *nomothetic* vs. *idiographic* approaches, first proposed by Windelband (1904) and discussed more thoroughly elsewhere (Allport, 1937; Sarbin, 1944) may be approached in one of its major aspects through the use of judgment models. If the argument is confined to that of "method of combination," as suggested by Meehl (1954), representational models are capable of providing some strikingly clear evidence. Martin (1957) presents some interesting findings in this respect, and it may be well for illustrative purposes to return to

Psychologist D, discussed above, and ask what the data suggest.

The first thing that might be said with respect to this psychologist is that, after having defined the predictors to his liking, he is not as astute as he might be in attaching subjective weights to them. By this it is meant that the discrepancies between subjective and relative weights are by no means insignificant in an absolute sense. The matter can be made even clearer by stating it another way: Given the variables in question, a computing machine would come closer to reproducing Ds judgments than he could come himself, were he to do the weighting as he says it should be done. What may be of particular importance is the fact that both of the interaction variables were overevaluated in this respect. Perhaps one likes to think of himself as being more complex than he actually is.

Since the computer does as well as it does in reproducing the judgments, an example is here provided wherein a single consistent set of rules of combination (nomothetic) may be successfully applied to the many and varied cases. The resulting judgments may well differ from "configural" or "clinical" judgments by amounts which are quite trivial. It is not impossible that this phenomenon will occur even in instances wherein the judge is schooled in the idiographic tradition and believes himself to be functioning accordingly. But this generalization and the earlier ones are perhaps reckless at this point, particularly in that the error variance for relative and subjective weights is not presented for these data. The purpose of the paper is more to illustrate the application of a methodology than to argue for reforms in clinical psychology from evidence based on a sample of Size 1.

What, additionally, may be said of the configurational model developed for D? We may legitimately ask to what extent the substitution of complex variables in the place of simple ones effectively enhanced the prediction of judgments, thereby testing the relative efficacy of two models. The R (corrected for shrinkage), using the configurational model, is .88. This appears high, and it may well be, except for the fact that the application of a linear model to D's judgments results in a corrected R of .91. Thus, substituting a configurational model for a linear one changes the proportion of predicted variance of judgments from 82.81% to 77.44%; a loss of a little better than 5%! And considering chance factors, the least that can be said is that there is no demonstrable gain over the linear model.

Summary

This paper has been concerned with the manner in which information is utilized in decision making or in judgment situations. It is shown that mathematical models provide a way of describing mental processes which would otherwise be accessible only through introspection or electrophysiological techniques. A linear model and a configurational model are described, and illustrations furnished for each. Such models make possible the testing of hypotheses concerning method of combination, individual differences in judgment ability, effects of training, personality correlates, idiographic interpretation of case materials, etc.

It may be said that the paramorphic representation of judgment appears to offer interesting possibilities for research and theory as well as for applied endeavors. Models which attempt to describe the method of combination of information in de-

cision making can illuminate problems which would otherwise remain obscure. In focusing upon the individual as the unit of research while at the same time preserving methodological rigor it becomes possible to achieve a level of psychological description which would otherwise be quite difficult. And few would disagree with the suggestion that sound description of the decision process is quite fundamental to a complete understanding of man.

REFERENCES

ALLPORT, G. W. *Personality, a psychological interpretation.* Holt, 1937.

BRUNSWICK, E. *Systematic and representative design of psychological experiments.* Berkeley: Univer. of California Press, 1947. (Also in J. Neyman [Ed.], *Berkeley symposium on mathematical statistics and probability.* Berkeley: Univer. California Press, 1949. Pp. 143–202.)

HAMMOND, K. R. Probabilistic functioning and the clinical method. *Psychol. Rev.*, 1955, **62**, 255–262.

HOLT, R. R. Clinical and statistical prediction: A reformulation and some new data. *J. abnorm. soc. Psychol.*, 1958, **56**, 1–12.

LUBIN, A. Some formulae for use with suppressor variables. *Educ. psychol. Measmt.*, 1957, **17**, 286–296.

LUFT, J. Implicit hypotheses and clinical predictions. *J. abnorm. soc. Psychol.*, 1950, **45**, 756–759.

MCNEMAR, Q. *Psychological statistics.* Wiley 1955.

MARTIN, H. T., JR. The nature of clinical judgment. Unpublished doctoral dissertation, Washington State Coll., 1957.

MEEHL, P. E. *Clinical vs. statistical prediction.* Minneapolis: Univer. Minnesota Press, 1954.

SARBIN, T. R. The logic of prediction in psychology. *Psychol. Rev.*, 1944, **51**, 210–228.

TODD, F. J. A methodological study of clinical judgment. Unpublished doctoral dissertation, Univer. of Colorado, 1954.

ULLMANN, L. P., & BERKMAN, VIRGINIA C. Judgments of outcome of home care placement from psychological material. *J. clin. Psychol.*, 1959, **15**, 28–31.

WINDELBAND, W. *Geschichte und Naturwissenschaft.* (3rd ed.) 1904.

(Received March 24, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 2, 1960

THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN:[1] REVIEW OF A DECADE OF RESEARCH

WILLIAM M. LITTELL
Claremont Graduate School

In the 10 years since its publication the Wechsler Intelligence Scale for Children (WISC) (Wechsler, 1949) has found wide acceptance among psychologists working with children in schools, clinics, and hospitals. With the wide use of the WISC, not only as a measure of intelligence but often as a clinical diagnostic instrument, it seems advisable to take a careful look at the growing fund of literature concerning the WISC, its validity in its various uses, and its general characteristics as a measuring device.

THE WISC

The WISC was developed as a downward extension of the Wechsler-Bellevue Intelligence Scales (W-B) (Wechsler, 1944), and most of the items contained in the WISC are from Form II of the adult scales (Wechsler, 1949; Seashore, Wesman & Doppolt, 1950). Easier items have been added to the low end of the subtests to make it suitable for use with young children.

Standardization. The WISC consists of 12 subtests grouped into a Verbal Scale (Information, Comprehension, Arithmetic, Similarities, Vocabulary, and Digit Span), and a Performance Scale (Picture Completion, Picture Arrangement, Block Design, Object Assembly, Coding, and Mazes). In the standardization of the WISC all 12 subtests were administered; only 10, however, were used to establish the IQ tables. Digit Span and Mazes were omitted "primarily [because of] their relatively low correlation with the other [sub] tests of the Scale and also, in the case of Mazes, the time factor" (Wechsler, 1949, p. 6). Wechsler suggests that all 12 subtests be given whenever possible "because of the qualitative and diagnostic data they add" (Wechsler, 1949, p. 6). When 11 or 12 subtests are used, prorating is necessary.

The WISC was standardized on 2200 white American boys and girls chosen to be representative of the 1940 census with respect to rural-urban residence, father's occupation, and geographic area. Some adjustment was made to allow for the "recent shift of population to the West." One hundred boys and 100 girls were chosen at each of 11 age levels, ages 5 through 15. Except for the 55 mentally deficient children included in the sample, all children were within one and one-half months of their mid-years. The mentally deficient group was drawn primarily from institutions in Illinois, Michigan, and New York, and more lenient age standards were observed. The standardization tests were administered by 17 field examiners who worked in 85 different communities.

The WISC IQ's (Verbal, Performance, and Full Scale) are deviation scores based on norms from other children of the same age. The raw scores obtained from subtests are transmuted into Scaled Scores by separate tables for each four-month age span (e.g., 5-0 to 5-3) and then

[1] The author wishes to express his appreciation to Robert Allen Keith for his many suggestions and comments through the several drafts of this paper.

into IQ's with a mean of 100 and a standard deviation of 15.

Early reviews of the WISC. Early reviews of the WISC varied from somewhat qualified acceptance (Delp, 1953b; McCandless, 1953) and the prediction of wide usage (Shaffer, 1949), to a rather critical rejection of the WISC in favor of the S-B (Anderson, 1953). All of the reviews were favorably impressed by the care taken in the standardization. Several specific criticisms were mentioned, however: the WISC manual lacks any evidence for its over-all validity (Delp, 1953b; McCandless, 1953; Shaffer, 1949); it provides a temptation to do elaborate pattern analyses on scores (McCandless, 1953) without providing substantiation for any interpretive value (Delp, 1953b; McCandless, 1953; Shaffer, 1949); it does not provide for extremely high (above 155) or extremely low (below 45) scores (Delp, 1953b); and the subtests appear to be too difficult for very young children (Delp, 1953b). The mental age was missed by Anderson (1953), Delp (1953b), and Shaffer (1949), and the lack of Negro children in the standardization group was mentioned as a weakness by McCandless (1953). Delp (1953b) felt further that the scoring of certain verbal items included considerable subjectivity. Anderson's rather critical review (1953) mentioned the fact that raw scores of zero are given scaled scores above zero for the younger children. Anderson was, in fact, able to find several apparent discrepancies in Wechsler's statistical treatment of the standardization data.

Strong points brought out by the reviewers of the WISC were its up-to-date construction (Delp, 1953b; McCandless, 1953) and its standardization (Anderson, 1953; Delp, 1953b; McCandless, 1953; Shaffer, 1949). Mentioned also as strong points were the facts that all of the children are administered comparable batteries (McCandless, 1953); the time of administration appears to be shorter (Delp, 1953b) and more predictable (McCandless, 1953) than comparable tests; it is easy to administer, interesting to children, gives both a Verbal and Performance IQ, provides IQ's directly comparable for various ages, appears to have potential clinical use, and has an easily used manual (Delp, 1953b).

A framework for evaluation. A word should be said concerning the framework within which this evaluation of the WISC is conducted. In addition to a number of articles reporting research on or with the WISC, the last 10 years have also shown advances in the methodology of psychological measurement and theory construction. (See especially Coombs, 1951; Coombs, Raiffa, & Thrall, 1954; American Psychological Association, 1954; Cronbach & Meehl, 1955.) The psychological test has come to be seen as only one element in the total process of theory construction. The full value of the test as a measure of a psychological variable depends upon how well the entire system in which it is used stands up to both logical and experimental test.

This view of the over-all "validity" of a test demands that (*a*) the area of the object world to be covered, (*b*) the nomothetic network containing the variable, and (*c*) the steps by which the test is demonstrated to be a measure of the variable, be made public, and that all assertions be subjected to empirical test.

This article is a review of the literature concerning the WISC since its publication in 1949. Its purposes are twofold: (*a*) to evaluate the WISC as a measure of various psy-

chological variables, and (*b*) to bring together for the user of the WISC information provided by the past decade of research.

The WISC as a Measure of Intelligence

Content Validity

In practice, a great deal of weight is often given to the user's assessment of the content validity of the WISC. The actual assessment is not simple, however, and is complicated by the similarity both in form and content between the WISC and the adult scales. As noted by others (Delp, 1953b; Shaffer, 1949), it tends to have attributed to it the validity of these other scales.

The content validity of a test involves a discussion of one's intuitive reasons for suspecting that a test will measure a certain variable, and is demonstrated " . . . by showing that the test items are a sample of a universe in which the investigator is interested" (Cronbach & Meehl, 1955, p. 282). In essence, the psychologist constructing a test involves content validity by including items which, in his opinion, will elicit behavior similar to the behavior he eventually hopes to predict. Said another way, if the test items are judged to be similar to the stimuli that ordinarily elicit the predicted behavior, he would then expect the test to demonstrate predictive validity. This predictive validity must be shown, however. To achieve any final evaluation of a test in terms of content validity alone is to rely upon intuition as a criterion, and, while intuition has a fine record as a guide for exploration, its use as a scientific criterion has consistently led to misinformation.

To assess the content validity of the WISC, a universe of items must be defined which is relevant to Wechsler's concept of children's intelligence. Unfortunately, beyond a few general remarks (Wechsler: 1949, 1950; Wechsler & Weider, 1953), no theoretical discussion of the concept of intelligence as it applies to children exists in print. To proceed, the assumption must be made that, at least in its more general aspects, the discussion of adult intelligence (Wechsler: 1944, 1958) is applicable to children.

Wechsler's definition of intelligence is very broad. As far as the trait "general intelligence" is concerned, any item which is judged to tap a child's "aggregate, or global capacity to act purposefully, to think rationally, and to deal effectively with his environment" (Wechsler, 1944, p. 3) might be included as a potential test item. Defined only at this rather gross level, it is difficult to conceive of any measure of directed behavior which would be definitely excluded.

A further assumption is made that a child's response to any intellective task is affected not only by his general intelligence, but by other "nonintellective" factors such as "drive" and "hand-eye coordination." While Wechsler presents rather convincing arguments for including such factors, the discussion of this universe is limited to a few examples. Wechsler states that he controlled for the differential effects of these "nonintellective" factors by including a wide variety of types of items (Wechsler, 1944).

As the test is constructed, two separate questions appear to be involved: (*a*) the sampling of the universe of relevant "nonintellective" factors by the different subtests (or combinations of subtests), and (*b*) the sampling of items within each subtest. By looking at explicitly stated theory, there seems to be no way in which the adequacy of the sampling of "nonintellective" factors

can be ascertained, for no statements are made to limit the possible range of factors.

On a less formal level, however, there seem to be several factors often included in any "common" concept of intelligence, but not adequately represented in the test. While these "omissions" would be of little consequence if the WISC were demonstrated to have the desired predictive validity, they might provide fruitful hypotheses if such validity is found to be lacking in any particular situation. Which test items, for instance, call for the integration of newly learned material into old contexts or for the memory of meaningful material? Further, the nature of the test situation rules out problem solving which takes place outside of a one-to-one relationship with another person or which involves any but very short periods of time.

The degree to which the items included within a given type represent an adequate sample for any particular child is a problem common to all intelligence tests, and presents another large source of question for the WISC. It is obvious, for example, that the degree to which a child would have a chance to learn the answer to "who wrote Romeo and Juliet?" or even "what is the color of rubies?" would differ markedly from one subculture to another. Yet success or failure on these items contributes equally to the IQ score no matter what the background of the child might have been. This type of criticism could also apply to subtests calling for specific skills such as putting puzzles together or manipulating a pencil.

In summary, the WISC appears to lack any explicitly stated, organized network of intuitive reasons for expecting it to show predictive validity other than the very broad assumption of a general factor which enters into the purposeful solution of all problems—*whether they occur in a test or in the child's life.* While Wechsler speaks convincingly of other, nonintellective factors which enter significantly into a child's actual behavior in problem situations, there appears to be little evidence that these factors are sampled in any systematic manner. This forces the user of the WISC to depend very heavily on whatever demonstrated criterion oriented and construct validity the WISC might have.

Predictive Validity

If the use of the term predictive validity is restricted to correlations between the WISC and some nontest measure of predicted behavior obtained at some time subsequent to the administration of the WISC, there are no relevant studies in the literature reviewed. This is very surprising, as it is difficult to conceive of any situation in which the WISC might be used that would not involve the prediction of behavior. As it stands, this lack of explicit evidence of the value of the WISC in the prediction of subsequent behavior must be viewed as a major weakness of the test.

Concurrent Validity

In general, reports of the concurrent validity of the WISC have been restricted to correlations between the WISC and other test measures of achievement or intelligence. Most of the studies relating the WISC to other intelligence tests have been oriented toward assessing the comparability of the various IQ scores.

Stanford-Binet. Studies reporting the comparability of WISC scores with S-B scores on different populations of children began to appear soon after the publication of the WISC.

A summary of these correlations appears in Table 1.

Frandsen and Higginson (1951) reported a study on 54 fourth-grade children and concluded that "IQ norms from the S-B and WISC are comparable at least within the range of one to two sigmas above and below the mean" (p. 283). This is the most favorable and unqualified statement of the comparability of the WISC and S-B appearing in the literature. An article by Pastovic and Guthrie (1951) followed summarizing the results of five unpublished master's theses. They concluded "that the WISC IQ should not be interpreted as equivalent to a Binet IQ at age levels below 10 years of age since the WISC score is consistently lower than that of the Binet" (p. 385).

Krugman, Justman, Wrightstone and Krugman (1951) found significant differences between the WISC Full Scale and Performance Scale IQs and the S-B IQ at all age levels

TABLE 1

STUDIES REPORTING CORRELATIONS BETWEEN THE WISC AND STANFORD-BINET, FORM L

Author	Subjects	*N*	Age Range	Correlations[a]		
		Boys Girls		V	P	FS
Frandsen and Higginson (1951)	4th grade children	54	9-1 to 10-3	.71	.63	.76
Krugman et al. (1951)	New York school children	332 166 166		.739	.644	.817
Nale (1951)	Mental defective children	104 54 50				.909
Sloan and Schneider (1951)	Mental defective children	40 20 20		.751	.641	.493
Stacey and Levin (1951)	Mental defective children	72		.69		.68
Weider et al. (1951)	White Louisville children—mean IQ below 90	44 23 21	5-0 to 7-11	.82	.79	.90
		62 38 24	8-0 to 11-11	.92	.78	.89
		Total	5-0 to 11-11	.89	.77	.89
Pastovic and Guthrie (1951)[b]	2nd grade children	50	7-6	.82	.71	.88
	Kindergarten children	50	5-6	.63	.56	.71
Clarke (1950)[b]	5th grade children	85	11-1	.83	.57	.79
Rapaport[b]	Public school children	100	7-6	.79	.74	.85
Cohen and Collier (1952)	Local Bloomington school children	51	Mean 7-5	.82	.80	.85
Mussen et al. (1952)	A "highly select population"	39	6-0 to 13-1	.83	.72	.85
Sandercock and Butler (1952)[c]	Mental defective children	90 58 32	10 to 16	.80	.66	.76
Triggs and Cartee (1953)	Mean S-B IQ 124.11	46	5-year-olds	.578	.478	.615
Arnold and Wagner (1955)	Elementary school children	50	8- & 9-year-olds	.88	.74	.90
Gehman and Matyas (1956)	School children	60	Mean 11-1	.78	.46	.73
	Same group—4 years later	29 31	Mean 15-2	.76	.64	.77
Stroud et al. (1957)	Children, Grades 3-6, referred to psychologist, mean IQ dull normal	621		.87	.83	.94
Schachter and Apgar (1958)	S-B administered 50.8 mo. before WISC	113 61 52	49.4 Mo.-S-B 100.2 Mo.-WISC	.64	.48	.67

[a] V = Verbal Scale; P = Performance Scale; FS = Full Scale.
[b] Studies summarized by Pastovic and Guthrie (1951).
[c] Study conducted using Form M.

(5–15), which were consistently in favor of the S-B. Differences between the S-B and WISC Verbal Scale tended to be significant only at younger age levels. They concluded further that "there is a definite tendency for greater differences . . . to be associated with the higher Stanford-Binet IQ's," and that differences between S-B and WISC Verbal and Full Scale IQs "tend to be associated with chronological age, in that such differences are larger at younger age levels" (p. 482).

It should be noted that a child cannot obtain an IQ above 154 on the WISC without extrapolation beyond the norms, while the S-B would allow much higher scores. This fact may explain in part the finding that the greater differences were associated with the higher S-B IQs.

Weider, Noller, and Schraumm (1951) also found that while the S-B and WISC IQs are significantly correlated, "the Binet IQ's tend to be higher than the WISC IQ's for the same children" (p. 332). A regression equation was suggested relating S-B to WISC Full Scale IQs in which WISC equals 0.85 Binet plus 11. According to this formula, when S-B IQs are below 73, the WISC IQs would be higher than the S-B IQs.

Cohen and Collier (1952), Mussen, Dean and Rosenberg (1952), and Stroud, Blommers and Lauber (1957) also reported correlations between the S-B and WISC. Further evidence that the WISC tends to score children within normal and upper ranges lower than the S-B is presented by Kureth, Muhr and Weisgerber (1952) in their study of 100 five- and six-year-old children, and by Levinson (1959) in his study of 117 Jewish preschool children. Triggs and Cartee (1953) tested 46 rather select children in the kindergarten of an independent school (S-B mean IQ of 124.11), and found WISC IQs to be consistently lower (Full Scale mean of 107.58). They concluded further that "there is a marked tendency for larger differences between Stanford-Binet and WISC IQ's to be related to higher Stanford-Binet IQ's" (p. 29).

Arnold and Wagner (1955) examined 50 children drawn at random from elementary schools and concluded that "so far as this sample is concerned, the relationship between IQ's obtained for eight- and nine-year-olds with the WISC (Full Scale) and Form L Binet is not significantly different from the relationship between IQ's obtained on Forms L and M of the Binet" (p. 93). The Verbal Scale related significantly better with the Binet than did the Performance Scale.

Preschool S-B IQs were compared with the school-age WISC IQs of 113 children selected at random from a clinic population born at a women's hospital (Schachter & Apgar, 1958). Of the 404 children requested by mail to return for testing, 119 returned for both tests; six were eliminated for other reasons. The resulting correlation of .67 (see Table 1) between the S-B and WISC Full Scale IQs was reported to compare favorably with previously reported correlations between preschool and school-age S-B IQs.

The comparability of IQ scores of the WISC and S-B when applied to mentally defective children has been investigated by several authors. Nale (1951) found the rather high correlation of .909 between the WISC Full Scale and the S-B, Form L, for 104 defective children, while Stacey and Levin (1951) and Sloan and Schneider (1951) report correlations of .68 and .493 respectively. In general, the WISC Full Scale was found to score somewhat higher than the S-B for these defective children.

Sandercock and Butler (1952) compared the WISC and S-B, Form M, IQs of 90 mentally defective children and concluded that "correlations obtained between the Stanford-Binet (M) and the three WISC IQs indicate a high degree of relationship between the Binet and WISC Verbal" (p. 104).

Several of the conclusions and assumptions made by various authors were subjected to direct test by Holland (1953) who found in part: (*a*) There was no significant practice effect on the WISC IQs when the S-B was given first and the median interval between the tests was seven days. (*b*) There was a significant difference between the correlations of the S-B with the Performance and with the Verbal and Full Scales of the WISC (in favor of the Verbal and Full Scales). (*c*) There was no significant difference between the correlations of the S-B with the Verbal and Full Scales of the WISC. (*d*) There was no significant relationship between chronological age and the difference between S-B and WISC IQs. (*e*) There was no significant relationship between S-B IQ and the difference between S-B and WISC IQs.

In general, the following conclusions can be drawn from these data about the comparability of the WISC and S-B IQs.

1. Studies involving a variety of ages and IQ ranges are very consistent in showing that at least within a white American school population the WISC and Stanford-Binet scores are related to a significant degree. Correlations between the WISC Full Scale and the S-B are predominantly reported within the .80's.

2. The WISC scores tend to be lower than S-B scores for the same children within the middle and upper ranges and somewhat higher for defectives. This appears to be particularly true for younger children (below 10) and for the higher S-B scores.

3. Using the S-B as a criterion, the highest correlations are found with the Full Scale IQ scores, the next highest with Verbal, and lowest with Performance scores.

Wechsler-Bellevue. The fact that the Wechsler-Bellevue (W-B) and the WISC overlap for the years 10 through 15 has led to several studies investigating the comparability of the WISC and W-B scores. The correlations reported are summarized in Table 2.

Knopf, Murfett, and Milstein (1954), feeling that the many similarities between the WISC and the W-B may suggest a comparability which is not actually there, administered the W-B and WISC to 30 Junior High School boys. They found that, while the WISC and W-B scores are highly correlated, the Verbal and

TABLE 2

STUDIES REPORTING CORRELATIONS BETWEEN THE WISC AND THE WECHSLER-BELLEVUE, FORM I

Author	Subjects	*N*	Age Range	Correlations[a]		
		Boys Girls		V	P	FS
Delattre and Cole (1952)	Public school children	50	10-5 to 15-7	.86	.82	.87
Vanderhost et al. (1953)	Mental defective children	38 22 16	11 to 16	.54	.77	.72
Knopf et al. (1954)	Jr. high school boys	30	13-4 to 14-6	.83	.64	.87

[a] V = Verbal Scale; P = Performance Scale; FS = Full Scale.

Full Scale scores on the WISC are significantly higher (at the .01 level). The Performance Scales, on the other hand, were not significantly different.

Price and Thorne (1955), testing two groups of white American public school children, found that at both the 11½- and 14½-year levels the WISC Full Scale and Verbal Scale IQ means tended to be higher than the corresponding W-B means, and that the direction of this difference was reversed for the Performance Scales. The authors set up criteria that two tests should be judged equivalent if, allowing for chance variation, (*a*) the individual should obtain essentially the same ranking on both tests, and (*b*) he should obtain essentially the same scores. By these criteria, at the 11½-year level the Verbal Scales were found to be lacking on both (*a*) and (*b*); the Performance Scales were found to be lacking on (*b*) and the Full Scales were remiss on neither. At the 14½-year level the Verbal and Full Scales were lacking on (*b*) and the Performance on (*a*).

Using as *S*s a group of 38 high-grade and borderline mental defectives, Vanderhost, Sloan and Bensberg (1953) also found the WISC Verbal Scale to score significantly higher than the W-B Verbal Scale, while no significant difference was found in Performance Scales. They concluded that because of this tendency for the W-B Verbal Scale to score significantly lower than the WISC Verbal Scale, the WISC is the preferred test to use on mental defectives in the 10- to 16-year range.

The following conclusions may be drawn about the comparability of the WISC and Wechsler-Bellevue in the age range over which they overlap.

1. The two scales appear to be related to a significant degree. Full Scale correlations are reported in the .70's and .80's.

2. The W-B Verbal Scale scores tend to be significantly lower than the WISC Verbal Scale scores for the same child. It may well be that the WISC items are more appropriate at this age level.

Other individual intelligence tests. In the following studies the WISC has often been used as the criterion against which the other test is validated. The results of these studies are reported in Table 3.

Three studies (Cohen & Collier, 1952; Pastovic & Guthrie, 1951; Sloan & Schneider, 1951) are reported in which the WISC has been correlated with the Arthur (see Table 3). The Arthur, as might be expected, appears to correlate better with the WISC Performance Scale than with the Verbal Scale.

Because of the length of time needed to administer the WISC, Martin and Wiechers (1954) investigated the possibility that the Colored Progressive Matrices could be used as a measure of intelligence with greater brevity than the WISC and a similar degree of validity. One hundred nine-year-old children from four Indiana schools were given the Matrices and the WISC in counterbalanced order. The authors concluded that "in view of these high correlations (see Table 3) and the ease and speed of administration it would seem that the Colored Progressive Matrices will find more extensive use in the clinical testing of children" (p. 144).

Following the positive results obtained by Martin and Wiechers, Stacey and Carleton (1955) investigated the degree to which the WISC and S-B scores of *S*s for a restricted range of intelligence (possible mentadefectives) compared with performl ance on the Colored Progressive

TABLE 3

STUDIES REPORTING CORRELATIONS BETWEEN THE WISC AND OTHER INTELLIGENCE TESTS

Author	Subjects	N (Boys / Girls)	Age Range	Test	Correlations[a] V	P	FS
McBrearty[b]	5th grade	52	11-2	Grace Arthur	.55	.65	.71
Sloan and Schneider (1951)	Mentally defective children	40 (20 / 20)		Grace Arthur, Form 1	.474	.833	.788
Cohen and Collier (1952)	Local Bloomington school children	47	7-5	Grace Arthur Revised, Form 2	.77	.81	.80
Martin and Wiechers (1954)	Indiana school children	100 (60 / 40)	9 years old	Colored Progressive Matrices	.84	.83	.91
Stacey and Carleton (1955)	Possible mental defective children	150 (80 / 70)	7-5 to 15-9	Colored Progressive Matrices	.54	.52	.55
Barratt (1956)	Entire 4th grade of a school	70 (26 / 34)	9-2 to 10-1	Progressive Matrices	.692	.699	.754
				Columbia Mental Maturity Scale	.559	.478	.606
Delp (1953a)	Public school children	74	6 to 15	Kent EGY	.596	.553	.618
Smith and Fillmore (1954)	Children with reading disability	91 (82 / 8		Ammons Full Range Picture Vocabulary Test	.73	.54	.75
Altus (1952)	Jr. high school children	55	13-7	Calif. Test of Mental Maturity (CTMM)			.81
Altus (1955)	Elementary children referred to guidance department	100		CTMM			
				Language	.71	.57	.70
				Non-language	.65	.67	.68
				Total	.76	.68	.77
Stempel (1953)	3rd and 4th grade children	50	8-5 to 10-4	SRA			
				Space	.49	.34	
				Number	.15	.38	
				Reasoning	.63	.55	
				Perception	.18	.42	
				Verbal Meaning	.67	.40	
				Total			.68
Cooper (1958)	5th grade children on Guam	51		Leiter International Performance Scale	.73	.78	.83
				Columbia Mental Maturity Scale	.66	.68	.74

[a] V = Verbal Scale; P = Performance Scale; FS = Full Scale.
[b] Study reported by Pastovic and Guthrie (1951).

Matrices. They found much lower correlations.

Motivated also by the time factor Barratt (1956) investigated the relationship between the WISC and the 1938 edition of the Progressive Matrices. Using 70 children who made up the entire fourth grade of a school, Barratt found correlations of .692, .699, and .754.

Because of the small number of studies reported, it is difficult to draw more than very tentative conclusions about the relation between the WISC and either form of the Progressive Matrices. It does appear, however, that when the Colored form is applied to a group of children with a normal spread of IQ scores, fairly high correlations can be expected, and that the Verbal and Performance Scales correlate equally well.

Investigating specifically the problem of testing children with reading difficulty, Smith and Fillmore (1954) reported a study correlating the WISC with the Ammons Full Range Picture Vocabulary Test, and concluded that as a screening device of intelligence the Ammons can be used with children with reading handicaps.

Delp (1953a), as part of a larger study, gathered data to compare the Kent Emergency Scales (EGY) with the WISC. He concluded that in view of the rather low correlations the primary value of the Kent EGY is not its correlation with the WISC, but its particular type of questions.

As part of a study to determine whether currently available tests would predict school achievement for bilingual pupils on the Territory of Guam, the WISC was administered to a sample of 51 fifth grade children (Cooper, 1958). In spite of the language handicap, significant correlations were reported with the Leiter International Performance Scale and the Columbia Mental Maturity Scale.

Group intelligence tests. Correlations between the WISC and the Science Research Associates Primary Mental Abilities Test are reported by Stemple (1953) and are shown in Table 4.

Altus (1952) reported correlations between the WISC and the California Test of Mental Maturity (CTMM). She selected a sample of 55 Junior High School children so as to represent the entire student body as to age, sex, proportion in each grade, proportion of bilinguals and IQ as measured by the CTMM. The correlation of .81 between the WISC Full Scale and the CTMM Total led her to conclude that "the WISC probably has considerable validity in comparable school settings" (p. 231).

A second study by Altus (1955), which was undertaken to test the assumption that the verbal and nonverbal portions of the WISC and the CTMM are significantly related, reported further correlations between these two tests (see Table 4). The 100 students referred to the guidance department by teachers included 36 who were referred for special training classes for the mentally retarded. Altus felt justified to conclude that "within a comparable school referral setting, the WISC and CTMM are markedly comparable as to group assessment and roughly comparable as to individual abilities."

While the three studies reviewed all report rather high correlations between the WISC and group intelligence tests, again the small number of studies precludes more than the very tentative acceptance of these conclusions.

Achievement tests. Mussen et al. (1952) reported a study with a group of Ohio State University elementary school children correlating WISC

TABLE 4

STUDIES REPORTING CORRELATIONS BETWEEN THE WISC AND MEASURES OF ACHIEVEMENT

Author	Subjects	N Boys	N Girls	Age Range	Measures of Achievement	Correlations[a] V	P	FS
Frandsen and Higginson (1951)	4th grade children	54		9-1 to 10-3	Stanford Achievement, Total	.62	.65	.76
Mussen et al. (1952)	A "highly select population"	21		6-0 to 13-1	Metropolitan Arithmetic	.74	.74	.81
					Reading	.62	.76	.75
		18			Stanford Arithmetic	.47	.29	.69
					Reading	.73	.57	.65
Sandercock and Butler (1952)	Mental defective children	58	32	10 to 16	"Achievement Quotient"	.53	.41	.51
Richardson and Surko (1956)	Delinquent children	90	15		Gray Oral Reading Paragraphs	.59		.58
		65			Stanford Achievement Test, Form D Reading			.59
					Arithmetic			.64
Barratt and Baumgarten (1957)	Achievers, Grades 4–6	30			California Achievement Test Reading	.61	.29	.56
					Arithmetic	.09	.14	.14
	Nonachievers, Grades 4–6	30			Reading	.51	.30	.61
					Arithmetic	.73	.33	.79
Stroud et al. (1957)	Referred for Psychological Study —Grades 3–6. Dull normal mean IQ	621			Iowa Test of Basic Skills Reading Comprehension	.58	.63	.66
					Arithmetic	.67	.52	.66
					Spelling	.62	.60	.67
Cooper (1958)	5th grade children on Guam	51			California Achievement Test	.80	.54	.77

[a] V = Verbal Scale; P = Performance Scale; FS = Full Scale.

scores with various measures of achievement. These correlations vary from .29 to .81. The fact that the intellectual range was limited by the "highly select population" may well have affected the obtained correlations adversely.

Frandsen and Higginson (1951) found rather consistent middle range correlations for fourth-grade children between the WISC scores and the Stanford Achievement Total score.

Barratt and Baumgarten (1957) related WISC scores to scores on the reading and arithmetic subtests of the California Achievement Tests for 30 achievers and 30 nonachievers in grades four to six. The achievers scored significantly higher on all scales of the WISC than the nonachievers. In both cases the Verbal Scales correlated higher with the reading subtest than did the Performance Scale. The almost chance relationship found between the WISC IQ's and the arithmetic achievement for achievers contrasted with the significant relationship between the two tests for nonachievers suggests strongly that other important variables are involved.

Sandercock and Butler (1952) found low positive correlations between a measure they call the Achievement Quotient and the WISC Scales for 90 mentally defective children. The Achievement Quotient was derived from judgments of the child's academic progress relative to his age. Further correlations with test measures of achievement for delinquent children were found by Richardson and Surko (1956).

Stroud et al. (1957) wished ". . . to determine the effectiveness with which all or various combinations of the WISC subtests could be used to predict performance on Reading Comprehension, Arithmetic, and Spelling tests of the Iowa Tests of Basic Skills battery" (p. 18). The tests were administered to 725 pupils in grades three to six drawn from a 21 county area in Iowa. All of the children had been referred for psychological interviews and testing and ". . . were in, or were thought to be in, some kind of school difficulty" (p. 18). The mean IQs were within the dull normal range. All of the various intercorrelations were calculated and beta weights for the various subtests determined. The authors found that the Arithmetic, Vocabulary, Block Design, and Object Assembly subtests were most effective in prediction for both the original group and a cross validation group of 129 like pupils. They concluded that their study gave no support for the use of separate verbal, nonverbal, and subtest scores in differential prediction.

The relation of the WISC IQ to another form of achievement was investigated by Robinowitz (1956) who wished to discover whether the brighter child as measured by an intelligence scale is the one who learns the relationship of opposition at an earlier age. Robinowitz found a significant difference (at the .01 level) in scores on the WISC between those children who were able to learn the relationship and those who were not. A point bi-serial correlation of .609 was found between the ability to learn the relation of opposition and scores on the WISC.

While not directly related to achievement, Mussen et al. (1952) reported correlations between teacher's rating of intelligence on the Haggerty-Olson-Wichman Rating Scale of Intelligence and the WISC of .64, .53, and .68 for the Verbal, Performance, and Full Scales respectively.

In general it would seem that the relationship between ability and achievement must be recognized as highly involved and complex, and should be subjected to much further investigation. At present it seems safe to say only that the WISC relates to scores on certain types of academic achievement tests for certain groups of children quite well. In general, the Verbal Scale seems to relate to test-measured academic achievement better than the Performance Scale.

Construct Validity

While an attempt at a full appraisal of the construct validity of the WISC would go far beyond the scope of this article, a few comments seem to be in order.

Concerning construct validity, Cronbach and Meehl (1955) state that "unless the network (nomethetic) makes contact with observations, and exhibits explicit, public steps of inference, construct validation cannot be claimed" (p. 291). At present, since little independent rationale exists for the WISC, it would seem that only a few rather general hypotheses could be drawn from the conceptual framework behind the WISC. In few studies is there an attempt to make these steps of inference explicit and public.

General intelligence. The assumption of the global nature of general intelligence is basic to the development of the Wechsler scales (Wechsler: 1944, 1949, 1958; Wechsler & Weider, 1953), and would imply that the WISC should correlate with other measures of general intelligence. The studies discussed under the heading of Concurrent Validity lend support to this view of general intelligence. It should be noted, however, that these studies lend support only to the assumption of a general trait which underlies all *test* behavior. The broader assumption of a general trait entering into *all* purposeful behavior both in and out of test situations is not touched by these studies.

Nonintellective factors. Also basic to Wechsler's theoretical position is the assumption that the particular subtests used in the WISC tap not only general intelligence, but other "nonintellective" factors. Some of these factors are specific to the particular subtest (e g., specific skills such as memory); others are more general and affect several or all of the subtests (e.g., "drive"). While these assumptions fit well into general testing theory in accounting for the various intercorrelations, it is very difficult to find any explicit statements about *which* subtests are affected by *what* other factors.

Both in discussion of the WISC and in its use a distinction is made between the Verbal and Performance Scales. Wechsler (1958) tentatively identifies the factors as measured by the *adult* scales as a verbal comprehension factor and a nonverbal organization factor (variously identified as performance, nonverbal, space, and visual-motor organization). Gault (1954) reported a factor analysis of the intercorrelations printed in the WISC Manual (Wechsler, 1949) and found the same general pattern of factors in the WISC as was reported by Hammer (1950) for the adult scales. The four factors worthy of note were called a "general eductive factor, a verbal comprehension factor, a spatial-perceptual factor and a memory factor" (p. 87). The verbal comprehension factor and the spatial-perceptual factor correspond roughly with the Verbal and Performance Scales.

Lotsof, Comrey, Bogartz, and

Arnsfield (1958) reported a factor analysis of WISC and Rorschach scores of 72 under-achieving children with reading disabilities. They found four factors which they called verbal intelligence, productivity, perceptual-movement, and performance speed. The Verbal and Performance Scales were not factorially pure, however; the Block Design was loaded significantly with the verbal intelligence factor, and Comprehension and Arithmetic were loaded with the performance speed factor. They concluded that "the verbal and performance aspects of the WISC are not independent of each other" (p. 301).

In general, evidence seems to support the rough factorial distinction between the Verbal and Performance Scales. Beyond this evidence on the division of the WISC into Performance and Verbal Scales, there seems to be no systematic investigation of the nature of any other of the somewhat general or specific factors tapped by the WISC subtests. This is of particular importance in evaluating the clinical use of the WISC and will be discussed in a later section.

Characteristics of the WISC as a Measuring Instrument

As with any measuring device, the user of an intelligence test must be familiar with the characteristics and idiosyncrasies of the test to be taken into account in any interpretation of the results. Several studies have been aimed either directly or indirectly at furnishing the WISC user with this information.

Reliability

Wechsler (1949) and Seashore et al. (1950) report coefficients of internal consistency (split-half reliabilities corrected by the Spearman-Brown formula) for all scales and for all subtests but Coding, Digit Span, and Mazes, at the 7½-, 10½-, and 13½-year levels. These figures range from .86 to .96. The coefficients of internal consistency for the various subtests range from .59 for Comprehension and Picture Completion at the 7½-year level to .91 for Vocabulary at the 10½-year level. The standard errors of measurement in IQ points for the three age levels for the Verbal Scale, Performance Scale, and Full Scale range from 3.00 to 5.61.

Both Wechsler (1949) and Seashore et al. (1950) warn the user to take into account the fairly low reliabilities of some of the subtests in interpreting either the absolute subtest scores or relations between them. For instance, at the 7½-year level only Vocabulary, Picture Arrangement, Block Design, and Mazes have coefficients of internal consistency above .70, while Comprehension and Picture Completion fall as low as .59. The reliability of the test in general tends to increase with age, so that at age level 13½ all subtests except Digit Span (.50) and Picture Completion (.68) are above .70.

The stability of the WISC scores over a four-year period has been investigated by Gehman and Matyas (1956). Sixty children were tested in the fifth grade and again in the ninth grade. Coefficients of stability for the three scales were: Verbal Scale, .77; Performance Scale, .74; and Full Scale, .77.

Sensitivity to Other Factors

Any measuring device, be it a surveyor's tape or an intelligence test, can be influenced by factors other than the ones the user wishes to measure. While WISC users appear to be aware of this fact, few studies appear which give direct information with which to evaluate any particular

WISC examiner-child interaction.

Practice effects. Holloway (1954), in an attempt to investigate the effect of a particular kindergarten program on the IQ scores of children, found that *both* his control and experimental groups showed significant gains (at the .01 level) in WISC Full Scale IQs over what appears from his report to be approximately a six-month period. The problem suggested by this study of the practice effects on repeated administrations of the WISC given over relatively short periods of time has not, to the writer's knowledge, been subjected to further direct investigation.

In studies in which the WISC and S-B or W-B have been administered in close temporal proximity, the authors have consistently reported no significant practice effects on the WISC scores (Kureth et al., 1952; Holland, 1953). It would not be safe to generalize from these findings to the WISC, however, for the case in which the test items are identical rather than more or less similar might well be different. This would seem especially true of performance items in which an "insightful" solution might be retained or of verbal items which might be taken back home or into the school room and discussed with others.

Variables in the test situation. The possible effects of differences in the examiner's technique of administration is another problem area which has not received the attention it merits, as is the whole field of possibilities arising from the relation between the examiner and the child and the circumstances of the examination. This is surprising, as the importance of these variables appears to be generally assumed.

Range of Application of the WISC

The literature provides considerable evidence that the WISC cannot be applied indiscriminately to all groups without considerable revision of the interpretation of the IQ score.

Southern Negro children. In connection with another study Young and Pitts (1951) tested 40 southern Negro children who were selected as a control group representative of their culture. These children were not retarded by socioeconomic criteria or by the judgment of observers. The mean WISC Full Scale IQ score of this group was, however, 69.8. To follow up on these results, Young and Bright (1954) tested a larger group of southern Negro rural children, and again found the markedly low mean WISC Full Scale IQ score of 67.74. The authors concluded that "We must question whether the WISC is a suitable test for the southern Negro child" (p. 220).

Bilingual children. Altus (1953) investigated the applicability of the WISC to children of bilingual Mexican descent. She compared the test patterning of these children with unilingual children equated for age, sex, and performance IQ and found that the Verbal Scales of the bilingual group were lower than the Performance Scales to a highly significant degree (a difference of nearly 17 points). No significant difference was found for the unilingual group. While this study was conducted with a group of children, the majority of whom had been referred for consideration for placement classes for mentally retarded, it again points out the need to exercise care in interpreting the IQs obtained from any markedly different group.

Levinson (1959) administered several intelligence tests to 117 Jewish preschool children and found that the S-B and all three WISC scores were higher at the .05 level of confidence

for unilingual children than for bilingual children.

Socioeconomic status. The possible effect of socioeconomic status was considered by Estes (1953) who administered the WISC to two groups of second- and fifth-grade children differing in socioeconomic status as measured by the Warner-Muhr-Eells Index of Status Characteristics. Significant differences in favor of the higher level were found in the total group of children and for the second-grade children. The difference for fifth-grade children was not significant. Levinson (1959), on the other hand, found no correlation between IQ level and socioeconomic background for Jewish preschool children.

Estes (1955) reported a follow-up of the earlier study in which 18 of the upper and 14 of the lower socioeconomic group were retested after a period of two years. The significant differences found when the children had been in the second grade no longer existed. The authors felt that this lessening of the effect of socioeconomic status reflected the increased "leveling" influence of the school with the passage of two years.

Laird (1957) tested two groups of 11-year old children differing in socioeconomic status but matched for a number of other variables. The mean score of the upper socioeconomic group fell within the bright normal range while the lower group had a mean score falling within the average range. Greater differences were found between Verbal and Full Scale scores than between Performance scores.

Mentally retarded. The largest single subgroup to which the WISC has been applied is the group of mentally retarded or deficient children. The question of the sensitivity of the WISC when used with these children was brought up by Carleton and Stacey (1955), who reported an item analysis of the WISC for "a sample of 366 subjects tested at Syracuse State School who can be classified as defective, borderline and dull normal" (p. 149). They found that for these children (*a*) relatively few items are misplaced with respect to order of presentation, and such misplacement as does occur does not seem to be of sufficient extent to affect materially the subtest total score, and (*b*) for each subtest there is a relatively abrupt shift from items which appear to be quite easy to ones which are quite difficult so that there are relatively few items of the middle range of difficulty.

A study by Stacey and Portnoy (1950) investigated the assumption that mental defective children will give responses to the WISC Vocabulary subtest at a lower conceptual level than borderline children. Two groups of children were tested (24 mental defective and 27 borderline) and their vocabulary responses were scored descriptive, functional, and categorical as representing increasing levels of concept formation. Contrary to expectation the borderline children gave significantly less functional and significantly more descriptive responses.

Deaf children. The possibility of using the WISC Performance Scale with deaf children was investigated by Graham and Shapiro (1953). Three groups of children were matched for physical health, sex, color, nativity, age and Goodenough Draw-a-Man IQ. Group (a) contained children with a 60 db or greater loss of hearing in both ears sustained prior to significant language development. The test had to be modified somewhat to make pantomime instructions possible. Groups (b) and (c) contained children with normal hearing. Each child was ad-

ministered the WISC Performance Scale; Groups (a) and (b) with pantomime instructions, and Group (c) with usual instructions. They found that Groups (a) and (b) did not differ significantly from each other, but were both significantly lower than Group (c).

The authors concluded that while the WISC Performance Scale cannot be used without modification as a valid measure of the intelligence of deaf children, it seems feasible to use a correction factor to nullify the effects of the pantomime instructions. In any case, they felt, the Performance Scale can be administered via pantomime as a crude measure.

Very young children. No studies are reported concerning the applicability of the WISC to the testing of very young children. It should be noted, however, that a child with a "mental age" of five or six or below would in effect be given subtests with as few as four or five items. The reliability of such short scales would be open to considerable question. In order to use the test at these ages, more items need to be added to the lower end of most scales. This criticism would, of course, apply to the use of the WISC with retarded children below the age of eight or nine years.

Summary:

1. There is strong evidence that WISC norms are not applicable to children of markedly different subgroups such as southern Negro and bilingual Mexican-American children.

2. Socioeconomic status appears to be a significant variable affecting the IQ scores of young children (second- as opposed to fifth-grade children), such that the children of higher socioeconomic status tend to obtain higher scores.

3. The WISC seems to be relatively insensitive to differences among mentally retarded children.

4. The WISC Performance Scale when administered with pantomime instructions to either normal or deaf children can be used as a crude and spuriously low measure of intelligence.

5. When the WISC is administered to children with "mental ages" below five or six years, the IQ scores can be expected to be relatively unreliable due to the limited number of "functional" test items at the low end of the scale.

Short Forms of the WISC

Two articles report attempts to develop short forms of the WISC. Carleton and Stacey (1954) made up 21 different short forms of the WISC from the WISC records of 365 children who had been referred to the Syracuse State School for evaluation and for whom there was no suspicion of organic involvement (IQ range 46 to 91). They correlated each of these short forms with the Full Scale IQ, finding correlations which ranged .64 for a two subtest combination (Comprehension and Vocabulary) to .88 for a five subtest combination of Comprehension, Arithmetic, Block Design, Coding, and Picture Completion.

Less hopeful results were reported by Yalowitz and Armstrong (1955), who derived three short form combinations from the WISC records of 229 children referred for numerous reasons to a child guidance clinic. Correlations with the Full Scale IQs ranged from .55 to .61. The authors felt that these low correlations may be attributed either to the "wide subtest scatter found in WISC records of

emotionally disturbed children, or . . . the lower subtest intercorrelations found in the WISC than on the Wechsler-Bellevue" (p. 277).

Armstrong (1955) divided the Vocabulary subtest of the WISC into two short forms consisting of odd and even words. The over-all split-half correlation for all ages five years no months to 14 years 11 months was .88. She concluded that "the loss of reliability involved in using either alternate word list instead of the total Vocabulary list is minimal, especially when compared to the time saved" (p. 414).

The Problem of Mental Age

The departure from the use of the concept of mental age has led both to comments and to suggested ways of finding an MA from the WISC scores. Grove (1950) felt that while the publication of the WISC was a real contribution, Wechsler had "thrown the baby out with the wash" when he discarded the concept of the MA along with its use as a "practical method of defining levels of test performance." The author then provided a method by which a mental age score could be obtained.

Wechsler (1951) himself, while still opposed to the MA as a measure of absolute intelligence, admitted that the MA concept has a use in comparing a child of a given age with children of his own age in performance on a given test. This test age, he felt, must be interpreted as a measure of "specific aptitude." He then outlined three different methods by which scores corresponding to "test age" can be calculated.

Kolstoe (1954) compared the performance of 29 third- and fourth-grade children (S-B IQ 116 or above) with 29 eighth- and ninth-grade children (S-B IQ 84 or below) on 11 of the 12 subtests of the WISC. Differences significant at the .05 level of significance were found on only three of the subtests. They concluded that their results "support to a considerable extent the generality of the mental age concept" (p. 167).

The WISC as a Diagnostic Instrument

In keeping with the growth of clinical psychology, tests previously used within a circumscribed area of prediction are finding use as more or less general diagnostic instruments. The WISC is, of course, a relatively standard sample of a child's behavior and, as such, can be used as any other "sample." Completely "disorganized" behavior, for instance, will have grossly similar diagnostic implications whether it occurs on the WISC, the Rorschach, or during a clinical interview. Beyond this use, however, there is a tendency to attempt to predict a wide variety of types of behavior from scores derived from the WISC.

Patterns of Subtest Deviations

As one might expect, the almost unlimited possibilities presented by 10 variables have engendered numerous hypotheses about how these variables relate to various aspects of a child's behavior. The problem of defining a "significant" deviation has been considered by Alimena (1951) who reported a method for achieving comparability of scores on the Wechsler subtests (for all Wechsler tests) and for evaluating their dispersion, based on the expected degree of trait variation within the individual. The author reported that the deviation norms have been calculated for the WISC and are available on request from him.

Differences between Verbal and Per-

formance scores. Recognizing that many WISC users tend to attribute meaning to any differences between a child's Verbal and Performance Scale scores, Seashore (1951) turned to Wechsler's original standardization data to investigate the meaning of such discrepancies. The WISC was originally designed so that the difference between average Verbal IQs and average Performance IQs was zero. Seashore found that the sigma of the difference scores for all ages was 12.5 and that the discrepancy scores closely approximate a normal distribution with mean 0.0. There were no important age differences in discrepancy scores.

Investigating the possible effects of group differences on the distribution of deviation scores, Seashore found no appreciable differences between rural and urban children, and that the feeble-minded group did not have a Performance score higher than their Verbal scores. Further, among the nine parental groups, only Professional and Semiprofessional showed any differences between mean Verbal IQ and mean Performance IQ. (Mean Verbal was about three points higher than mean Performance for both groups.)

Newman and Loos (1955) investigated specifically whether there are differences between the Verbal IQ scores and Performance IQ scores for mentally defective children. They found that (*a*) mentally defective children classed as familial ($N = 128$) obtained significantly higher scores on the Performance tests than on the Verbal tests (mean difference was 8.07), (*b*) mentally defective children classed as undifferentiated ($N = 75$) also performed significantly higher on the Performance than on the Verbal tests, but to a lesser degree than the familial (mean difference was 4.8), (*c*) mentally defective children due to brain damage or birth trauma and giving no evidence of severe motor defect showed no difference, and (*d*) the brain-damaged children showed significantly lower Performance scores than the undifferentiated group.

Both Sloan and Schneider (1951) and Stacey and Levin (1951) also found the Performance Scale to score significantly higher than the Verbal Scale for the mentally deficient children they examined. In general, it seems that one should expect mentally retarded children classified as familial or undifferentiated to obtain higher Performance than Verbal Scale scores.

On the other hand, Atchison (1955) found that the 80 feeble-minded Negro boys and girls he tested tended to score higher on the Verbal Scale than on the Performance Scale, reversing the differences found above. It would seem safe to assume that there are important variables involved in the relationship between Verbal and Performance Scale scores which were not controlled adequately in the above studies.

Application of Hypotheses from the W-B. Hypotheses abound concerning patterns of deviations on the W-B and Delattre and Cole (1952) were concerned lest psychologists might attempt to use these cues in interpreting the WISC. Consequently, they compared the profiles of 50 W-B, Form I, protocols with the patterns obtained from WISCs of the same children. The data were analyzed to determine the extent to which the relative position of a subtest to the scaled mean occurring on the one test was likely to be repeated on the other. They concluded that the similarity of profiles is not large enough to warrant prediction in indi-

vidual cases, and, while the IQs will tend to be grossly similar, the clinical sign approach cannot be carried over from the W-B to the WISC.

It should be noted that Rabin and Guertin (1951) in their review of the W-B through 1950 conclude that "the scatter mountain gave birth to a mouse" (p. 240). The numerous studies they review suggested that ". . . the various measures of scatter and variability—the different patterns have succeeded in differentiating [some] *groups*, but not *individuals*" (p. 240).

Reading difficulty. The question of a WISC pattern for children with marked reading difficulties has caught the attention of several authors. Altus (1956) reported finding a distinctive test pattern for children with severe reading disabilities. The records of 25 children (24 boys and 1 girl) who showed a discrepancy of two years or more between their expected and actual reading level were investigated. Coding and Arithmetic subtests were found to be significantly lower than Vocabulary, Digit Span, Picture Completion, Object Assembly, and Picture Arrangement at the .01 level of significance; Information was lower than Picture Completion at the .01 level and lower than Digit Span at the .02 level. Altus found that these results were quite similar to W-B results on illiterate soldiers, but did not state her criteria for similarity.

In an intelligence test of 10 subtests, the chances that at least one subtest would deviate significantly from the mean of *all* of the others at the .01 level is one in 10. This factor takes on particular importance in the above study, for there was no rationale stated prior to the study by which one would expect any particular test to deviate.

Graham (1952) wished to check the assertion by others that the W-B and WISC profiles of unsuccessful readers and psychopathic adolescents are similar. He went over the records of 96 unsuccessful readers (25% or more below the mean of the Wide Range Achievement Test) who had achieved either a Verbal or Performance Scale score of 90 or above, comparing the mean scatters with the previously obtained (but unpublished) scatter of adolescent psychopaths. Graham reported no statistics but concluded that these patterns "correspond closely." For the unsuccessful reader, he found Arithmetic, Digit Span, Information, Digit Symbol, and Vocabulary subtest averages below the mean, and Object Assembly, Picture Completion, Picture Arrangement, Block Design, Comprehension, and Similarities subtest averages above the mean. Only Arithmetic and Similarities deviated to a degree significant at the .01 level.

A comparison of these results with those of Altus (1956) finds Arithmetic to be significantly lower than the others in both studies. Of the other subtests mentioned in both studies six of the subtests deviate in similar directions while two deviate in opposite directions.

Organic brain damage. One study concerns itself with finding subtest patterns characteristic of children with organic brain damage. Beck and Lam (1955) investigated the WISC records of 104 children referred as possible candidates for a special class for the educable mentally retarded. These children were placed into three groups: (a) organic ($N=27$), diagnosed by neurological examination; (b) suspected organics ($N=48$), inferred by psychological studies; and (c) non-organic ($N=29$),

for whom there was no evidence of organicity from psychological evaluation or developmental history. Eleven more children were added to Group (c) a year later. From a comparison of the mean Verbal, Performance, and Full Scale scores and of the intersubtest scatter, he concluded that (*a*) organics tend to score lower on the WISC Full Scale than nonorganics, (*b*) organics tend to score lower on the WISC Performance and Full Scales than on the Verbal Scale, (*c*) the possibility of organic damage increases considerably as the IQ drops below the 70–80 range, and (*d*) the WISC does not show a characteristic pattern of subtest scores for organics as a group (as opposed to nonorganic, possibly mentally retarded children).

The Interpretation of Individual Subtest Scores

As noted above, Wechsler (1944, 1958) assumes that specific subtests tap not only general intelligence, but specific factors as well. The exact nature of these factors, however, is far from clear. Some hints are given by Wechsler (1944, 1958) as to what he considers these factors to be for the adult scales; no help is given in interpreting the meaning of the subtests of the WISC when applied to children, however, beyond the statement that the subtests seem to measure different factors in children than in adults (Wechsler, 1949). Balinsky (1941) found evidence to suggest that even within the adult scales the subtests do not measure the same factors at all age levels.

Nowhere in the literature covered is there more than the barest beginning of the investigation of the various interpretive hypotheses. It would appear that most, if not all, are based on an intuitive appraisal of the content of the subtest and the informal observations of test administrators. While some agreement might be found as to the most likely interpretation of some subtest scores (e.g., Digit Span), other subtests (e.g., Similarities) might produce wide disagreement. Even if one could find agreement as to what a particular item should measure, the question of empirical validation would remain. It should be noted further that most, if not all, of the coefficients of internal consistency would cast much doubt on any individual prediction. In the last analysis, it would seem that any prediction made on the basis of an individual subtest score is little more than a rationalized hunch. A plausible rationale certainly does not make a valid measure.

Summary

This article has reviewed the literature concerning the WISC since its publication in 1949. Its purposes were twofold: (*a*) to evaluate the WISC as a measure of various psychological variables, and (*b*) to bring together for the user of the WISC information brought out by the past decade of research. The WISC has been discussed in terms of its validity as a measure of intelligence, its characteristics as psychological measuring device, and its use as a diagnostic instrument.

As summaries have been provided whenever appropriate within the body of the article, no attempt will be made here to repeat all of the points considered or information brought out. A few general statements do, however, seem in order concerning some rather important areas of unmet need.

Aside, perhaps, from correlations between the WISC and Stanford-Binet for normal white school children, further investigation of any of the problems discussed could add significantly to our fund of knowledge, both practical and theoretical, concerning the WISC and its use. Three areas in particular, however, stand out.

1. The WISC does not have an adequate rationale. Much more thought and effort need to be devoted to putting the WISC on a firm theoretical foundation. At present, both the assessment of the test's content validity and the long process of construct validation are severely handicapped by this lack of an explicit rationale.

2. The lack of investigations of the test's predictive validity in its many common uses is appalling. At present, the test's content and construct validities are not strong enough to support the use of the test without this criterion-oriented validation. It would seem that all possible occasions should be taken to discover experimentally, if the WISC does indeed predict what it is assumed to predict. For example, children are placed in classes for the mentally retarded on the assumption that they will respond to various learning situations in characteristic ways. How well does the WISC predict this response in a well-controlled, experimental situation?

3. Much more systematic attention should be given to investigations of the many practical problems involved in the use of the WISC as a measuring device. There appears to be strong reason to suspect that WISC scores are affected systematically by many variables other than intelligence, but little information about the exact nature of these variables and the relationships involved is available. Especially in need of systematic investigation is the effect on WISC scores of (*a*) variables in the relationship between examiner and examinee, (*b*) the circumstances of the examination, and (c) repeated administrations of the WISC.

On the other hand, the WISC appears to be a relatively well-standardized test with many virtues. It correlates consistently well with other measures of intelligence, appears to be widely accepted and used, and, in general, seems to merit further research and development.

REFERENCES

ALIMENA, B. Norms for scatter analysis on the Wechsler Intelligence Scales. *J. Clin. Psychol.*, 1951, **7**, 289–290.

ALTUS, GRACE T. A note on the validity of the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1952, **16**, 231.

ALTUS, GRACE T. W.I.S.C. patterns of a selective sample of bilingual school children. *J. genet. Psychol.*, 1953, **83**, 241–248.

ALTUS, GRACE T. Relationship between verbal and non-verbal parts of the CTMM and WISC. *J. consult. Psychol.*, 1955, **19**, 143–144.

ALTUS, GRACE T. A WISC profile for retarded readers. *J. consult. Psychol.*, 1956, **20**, 155–156.

AMERICAN PSYCHOLOGICAL ASSOCIATION, Committee on Psychological Tests. *Technical recommendations for psychological tests and diagnostic technique.* Wash. D. C.: APA, 1954.

ANDERSON, J. M. Review of the WISC. In O. K. Buros (Ed.), *Fourth ment. Measmt. Yearb.*, Highland Park, New Jersey: Gryphon, 1953. Pp. 480–481.

ARMSTRONG, RENATE G. A reliability study of a short form of the WISC vocabulary subtest. *J. clin. Psychol.*, 1955, **11**, 413–414.

ARNOLD, F. C., & WAGNER, WINIFRED K. A comparison of Wechsler Children's Scale and Stanford-Binet scores for eight- and

nine-year olds. *J. exp. Educ.*, 1955, **24**, 91–94.

ATCHISON, C. O. Use of the Wechsler Intelligence Scale for Children with eighty mentally defective Negro children. *Amer. J. ment. Defic.*, 1955, **60**, 378–379.

BALINSKY, B. An analysis of mental factors of various age groups from nine to sixty. *Genet. Psychol. Monogr.*, 1941, **23**, 191–234.

BARRATT, E. S. The relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC. *J. consult. Psychol.*, 1956, **20**, 294–296.

BARRATT, E. S., & BAUMGARTEN, DORIS L. The relationship of the WISC and Stanford-Binet to school achievement. *J. consult. Psychol.*, 1957, **21**, 144.

BECK, H. S., & LAM, R. L. Use of the WISC in predicting organicity. *J. clin. Psychol.*, 1955, **11**, 154–157.

CARLETON, F. O., & STACEY, C. L. Evaluation of selected short forms of the Wechsler Intelligence Scale for Children. *J. clin. Psychol.*, 1954, **10**, 258–261.

CARLETON, F. O., & STACEY, C. L. An item analysis of the Wechsler Intelligence Scale for Children. *J. clin. Psychol.*, 1955, **11**, 149–154.

COHEN, B. D., & COLLIER, MARY J. A note on WISC and other tests of children six to eight years old. *J. consult. Psychol.*, 1952, **16**, 226–227.

COOMBS, C. H. A theory of psychological scaling. *Univer. Mich. Engng. Res. Inst. Bull.*, 1951, No. 34.

COOMBS, C. H., RAIFFA, H., & THRALL, R. M. Some views on mathematical models and measurement theory. In R. M. Thrall, C. H. Coombs, & R. L. Davis (Eds.), *Decision processes*. New York: Wiley, 1954. Pp. 19–37.

COOPER, J. G. Predicting school achievement for bilingual pupils. *J. educ. Psychol.*, 1958, **49**, 31–36.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

DELATTRE, LOIS, & COLE, D. A comparison of the WISC and the Wechsler-Bellevue. *J. consult. Psychol.*, 1952, **16**, 228–230.

DELP, H. A. Correlations between the Kent EGY and the Wechsler batteries. *J. clin. Psychol.*, 1953, **9**, 73–75. (a)

DELP, H. A. Review of the WISC. In O. K. Buros (Ed.), *Fourth ment. Measmt. Yearb.*, Highland Park, New Jersey: Gryphon, 1953. (b)

ESTES, BETSY W. Influence of socioeconomic status on Wechsler Intelligence Scale for Children: An exploratory study. *J. consult. Psychol.*, 1953, **17**, 58–62.

ESTES, BETSY W. Influence of socioeconomic status on Wechsler Intelligence Scale for Children: Addendum. *J. consult. Psychol.*, 1955, **19**, 225–226.

FRANDSEN, ARDEN N., & HIGGINSON, J. B. The Stanford-Binet and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 236–238.

GAULT, UNA. Factorial patterns on the Wechsler Intelligence Scales. *Aust. J. Psychol.*, 1954, **6**, 85–90.

GEHMAN, ILA H., & MATYAS, R. P. Stability of the WISC and Binet tests. *J. consult. Psychol.*, 1956, **20**, 150–152.

GRAHAM, E. E. Wechsler-Bellevue and WISC scattergrams of unsuccessful readers. *J. consult. Psychol.*, 1952, **16**, 268–271.

GRAHAM, E. E., & SHAPIRO, ESTHER. Use of the Performance Scale of the Wechsler Intelligence Scale for Children with the deaf child. *J. consult. Psychol.*, 1953, **17**, 396–398.

GROVE, W. R. Mental age scores for the Wechsler Intelligence Scale for Children. *J. clin. Psychol.*, 1950, **6**, 393–397.

HAMMER, A. G. A factor analysis of Bellevue tests. *Aust. J. Psychol.*, 1950, **1**, 108–114.

HOLLAND, G. A. A comparison of the WISC and Stanford-Binet IQ's of normal children. *J. consult. Psychol.*, 1953, **17**, 147–152.

HOLLOWAY, H. D. Effects of training on the SRA Primary Mental Abilities (Primary) and the WISC. *Child Develpm.*, 1954, **25**, 254–263.

KNOPF, I. J., MURFETT, BETTY J., & MILSTEIN, V. Relationships between the Wechsler-Bellevue Form I and the WISC. *J. clin. Psychol.*, 1954, **10**, 261–263.

KOLSTOE, O. P. A comparison of mental abilities of bright and dull children of comparable mental ages. *J. educ. Psychol.*, 1954, **45**, 161–168.

KRUGMAN, JUDITH I., JUSTMAN, J., WRIGHTSTONE, J. W., & KRUGMAN, M. Pupil functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 475–483.

KURETH, GENEVIEVE, MUHR, JEAN P., & WEISGERBER, C. A. Some data on the validity of the Wechsler Intelligence Scale for Children. *Child Develpm.*, 1952, **23**, 281–287.

LAIRD, DOROTHY S. The performance of two groups of eleven-year-old boys on the Wechsler Intelligence Scale for Children. *J. educ. Res.*, 1957, **51**, 101–108.

LEVINSON, B. M. A comparison of the performance of bilingual and monolingual native born Jewish preschool children of traditional parentage on four intelligence tests. *J. clin. Psychol.*, 1959, **15**, 74–76.

LOTSOF, E. J., COMREY, A., BOGARTZ, W., & ARNSFIELD, P. A factor analysis of the WISC and Rorschach. *J. proj. Tech.*, 1958, **22**, 297–301.

MCCANDLESS, B. R. Review of the WISC. In O. K. Buros (Ed.), *Fourth ment. Measmt. Yearb.*, Highland Park, New Jersey: Gryphon, 1953. Pp. 480–481.

MARTIN, A. W., & WIECHERS, J. E. Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1954, **18**, 143–144.

MUSSEN, P., DEAN, S., & ROSENBERG, MARGERY. Some further evidence on the validity of the WISC. *J. consult. Psychol.*, 1952, **16**, 410–411.

NALE, S. The Childrens-Wechsler and the Binet on 104 mental defectives at the Polk State School. *Amer. J. ment. Defic.*, 1951, **56**, 419–423.

NEWMAN, J. R., & LOOS, F. M. Differences between Verbal and Performance IQ's with mentally defective children on the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1955, **19**, 16.

PASTOVIC, J. J., & GUTHRIE, G. M. Some evidence on the validity of the WISC. *J. consult. Psychol.*, 1951, **15**, 385–386.

PRICE, J. R., & THORNE, G. D. A statistical comparison of the WISC and Wechsler-Bellevue, Form I. *J. consult. Psychol.*, 1955, **19**, 479–482.

RABIN, A. I., & GUERTIN, W. H. Research with the Wechsler-Bellevue Test: 1945–1950. *Psychol. Bull.*, 1951, **48**, 211–248.

RICHARDSON, HELEN M., & SURKO, ELISE F. WISC scores and status in reading and arithmetic of delinquent children. *J. genet. Psychol.*, 1956, **89**, 251–262.

ROBINOWITZ, R. Learning the relation of opposition as related to scores on the Wechsler Intelligence Scale for Children. *J. genet. Psychol.*, 1956, **88**, 25–30.

SANDERCOCK, MARIAN G., & BUTLER, A. J. An analysis of the performance of mental defectives on the Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1952, **57**, 100–105.

SCHACHTER, FRANCES F., & APGAR, VIRGINIA. Comparison of preschool Stanford-Binet and school-age WISC IQS *J. educ. Psychol.*, 1958, **49**, 320–323.

SEASHORE, H. G. Differences between Verbal and Performance IQ's on the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 62–67.

SEASHORE, H., WESMAN, A., & DOPPOLT, J. The standardization of the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1950, **14**, 99–110.

SHAFFER, L. F. Review of the WISC. *J. consult. Psychol.*, 1949, **13**, 453–454.

SLOAN, W., & SCHNEIDER, B. A study of the Wechsler Intelligence Scale for Children with mental defectives. *Amer. J. ment. Defic.*, 1951, **55**, 573–575.

SMITH, L. M., & FILLMORE, ARLINE R. The Ammons FRPV Test and the WISC for remedial reading cases. *J. consult. Psychol.*, 1954, **18**, 332.

STACEY, C. L., & CARLETON, F. O. The relationship between Raven's Colored Progressive Matrices and two tests of general intelligence. *J. clin. Psychol.*, 1955, **11**, 84–85.

STACEY, C. L. & LEVIN, JANICE. Correlation analysis of scores of subnormal subjects on the Stanford-Binet and Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1951, **55**, 590–597.

STACEY, C. L., & PORTNOY, B. A study of the differential responses on the vocabulary sub-test of the Wechsler Intelligence Scale for Children. *J. clin. Psychol.*, 1950, **6**, 401–403.

STEMPEL, ELLEN F. The WISC and the SRA Primary Mental Abilities Test. *Child Develpm.*, 1953, **24**, 257–261.

STROUD, J. B., BLOMMERS, P., & LAUBER, MARGARET. Correlation of WISC and achievement tests. *J. educ. Psychol.*, 1957, **48**, 18–26.

TRIGGS, F. O., & CARTEE, J. K. Pre-school pupil performance on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *J. clin. Psychol.*, 1953, **9**, 27–29.

VANDERHOST, LEONETTE, SLOAN, W., & BENSBERG, G. J., JR. Performance of mental defectives on the Wechsler-Bellevue and the WISC. *Amer. J. ment. Defic.*, 1953, **57**, 481–48[illegible]

WECHSLER, [illegible] *Measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.

WECHSLER, D. *Wechsler Intelligence Scale for Children*. New York: Psychological Corp., 1949.

WECHSLER, D. Intellectual development and psychological maturity. *Child Develpm.*, 1950, **21**, 44–50.

WECHSLER, D. Equivalent test and mental ages for the WISC. *J. consult. Psychol.*, 1951, **15**, 381–384.

WECHSLER, D. *Measurement and appraisal of adult intelligence.* Baltimore: Williams & Wilkins, 1958

WECHSLER, D., & WEIDER, A. Tests of intelligence. C. Wechsler Intelligence Scale for Children. In A. Weider (Ed.), *Contributions toward medical psychology.* New York: Ronald Press, 1953. Pp. 522–529.

WEIDER, A., NOLLER, P. A., & SCHRAUMM, T. A. The Wechsler Intelligence Scale for Children and the Revised Stanford-Binet. *J. consult. Psychol.*, 1951, **15**, 330–333.

YALOWITZ, J. M., & ARMSTRONG, RENATE G. Validity of short forms of the Wechsler Intelligence Scale for Children (WISC). *J. clin. Psychol.*, 1955, **11**, 275–277.

YOUNG, FLORENCE M., & BRIGHT, H. H. Results of testing 81 Negro rural juveniles with the Wechsler Intelligence Scale for Children *J. Soc. Psychol.*, 1954, **39**, 219–226.

YOUNG, FLORENCE M., & PITTS, VIRGINIA A. The performance of congenital syphilitics on the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 239–242.

(Received April 4, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 2, 1960

COMMENTS ON "INTRACLASS CORRELATION VS. FACTOR ANALYTIC TECHNIQUES FOR DETERMINING GROUPS OF PROFILES"

HAROLD P. BECHTOLDT
State University of Iowa

The purpose of this discussion is to analyze briefly a few statements presented in a paper by Haggard, Chapman, Isaacs, and Dickman (1959) entitled "Intraclass correlation vs. factor analytic techniques for determining groups of profiles." They are concerned chiefly with comparing the "factor analytic" and "direct correlation" methods as two approaches for grouping profiles, a problem in typology. Careful reading of this paper reveals a number of ambiguous or incorrect presentations of technical and procedural matters. If these technical matters are not clarified, other investigators interested in applying these two approaches to problems of profile comparisons may be seriously misled. The present discussion will not consider the more fundamental question of the usefulness, for any scientific purpose, of any method of grouping individuals on the basis of the "shape," "level," and "scatter" attributes of a set of numbers termed a profile. (Cronbach & Gleser, 1953)

FACTOR ANALYTIC TECHNIQUES

Readers of the paper by Haggard et al. may well gain the incorrect impression that the orthogonal centroid method, the centroid method with an oblimax solution, and the multiple group method are three, or at least two, different methods of factor analysis. The early misconception of the multiple group and centroid techniques as distinct methods has been clarified in papers by Guttman (1944, 1952) and Harman (1954). The multiple group procedure is simply a computing technique which may be used to define one, but usually simultaneously two, or more composite variables or "factors" from a set of observations. Appropriate application of the multiple group procedures will lead to any of the possible orthogonal or oblique axes or "factors." The necessary sets of weights for defining the desired "factors" may be obtaining in a number of ways; one may use any of the several "complete centroid" or "group centroid" methods with or without "rotational" transformations, or one might apply cluster methods based on an inspection of the data. An investigator may use some a priori "theoretical" or even (within limits) an arbitrary formulation for the weights or supplemental information such as clinical ratings of cases as "hypertensive," "neurotic," or "psychotic," the procedure used in the paper under discussion.

The "oblimax rotational solution" is simply one of the several available analytical procedures, each of which defines the "factors" by an objective transformation of certain types of data called "factor loadings." As defined by Thurstone (1947), factor loadings are the "orthogonal projections" of the variates, considered as vectors, upon a set of special axes added to the system and named "normals" or "reference axes." The oblimax procedure is a method for "rotating" such a set of reference axes, either orthogonal or oblique, to positions defined by a mathematically

stated criterion.

The common misconception that the oblimax procedure[1] is restricted to rotating orthogonal reference axes may account in part for the inappropriate comparison by Haggard et al. (1959) of the oblimax data (Set A) with the data from the multiple group method (Set B) shown in their Table 2, page 51. However, their discussion of these two sets of data in terms of factor loadings and projections as well as their untenable distinction between "oblique space" and "orthogonal space" suggest a lack of understanding of the basic concepts of factor analysis as developed by Thurstone (1947) and by Holzinger and Harman (1941).

The computations for the Set B data and Set C data have been checked by the writer and have been found to be accurate to within $\pm .02$. Both of these sets of data represent *orthogonal* projections on a set of "primary axes" defined by the centroids of three "clusters" or groups of profiles; such projections are named the "factor structure" by Holzinger and Harman (1941), although Haggard et al. refer to the Set B values as the "factor pattern" and use the term "factor structure" to refer to the intercorrelations among the factors (p. 52). The Set B and Set C data are derived from two somewhat different definitions of "clusters" of profile vectors.

The Set A values are entitled the "centroid solution, oblimax rotation." This designation of the Set A data is ambiguous. The Set A values from an oblimax program might be expected to be either *orthogonal* projections on a set of "normals," or *oblique* projections on a set of "primary axes," these oblique projections being referred to by Holzinger and Harman as the "factor pattern." However, the values could also represent two other sets of projections. Because of what apparently are computational errors in the Set A data in addition, an unambiguous identification of these values could not be made without repeating the entire set of calculations.

Unfortunately, even when all the computations are correct, the values represented by the Set A data and by the Set B data probably are not comparable sets of values for an oblique system. The oblimax data either are linear regression coefficients expressing the variates in terms of the factors when the values are oblique projections on the primaries or, as "factor loadings," are proportional to regression coefficients. The Set B data are covariances between the factors and the (profile) variables. Covariances and regression coefficients are not generally comparable numerically. However, the regression coefficients and covariances for any one set of defined factors and observed variates are related by a simple equation (Holzinger & Harman, 1941, p. 327). Contrary to the implications of statements in the paper under discussion, the values in the Table 3, page 51, are not relevant to this problem.

In a factor analysis, the distinctions between reference axes and primary axes on the one hand and between orthogonal and oblique projections on the other can be very important. The fairly obvious differences in numerical values between the three sets of data labeled Set B, Set E, and Set F in Table 1 below provide an illustration of these distinctions. Each of these three sets of data or "factor matrices" and even a fourth matrix of "reference vector pattern values" might be referred to

[1] The use of the oblimax solution with any set of reference axes was clarified by Kern Dickman.

TABLE 1

THREE ALTERNATIVE METHODS FOR REPRESENTING A GIVEN FACTOR ANALYSIS SOLUTION OBTAINED WITHOUT ROTATION

	Set B			Set E			Set F			
	Orthogonal Projections on Cluster Primary Axes (Factor Structure)			Oblique Projections on Cluster Primary Axes (Factor Pattern)			Orthogonal Projections on Normals or Reference Axes (Factor Loadings)			h^2 Communality Estimates
	I	II	III	I	II	III	I	II	III	
1	88	49	−74[a]	63	14	−25	36	12	−16	.801
2	87	18	−70	101	−32	−02	58	−27	−01	.840
3	101	68	−59	110	19	25	63	16	16	1.078
4	56	100	−21	15	95	08	09	79	05	1.008
5	26	96	07	−14	105	16	−08	87	10	.990
6	44	77	−33	−20	81	−33	−11	67	−21	.644
7	51	88	−19	18	82	09	10	68	06	.801
8	−40	09	82	36	11	110	21	09	71	.745
9	−83	−22	99	−27	06	81	−15	05	52	1.007
10	−84	−37	96	−18	−13	81	−10	−11	52	.983
11	−62	05	96	02	23	102	01	19	66	.973
12	−72	−40	90	07	−27	90	04	−22	58	.850

Note.—Data computed from 12 MMPI profiles presented by Haggard et al. (1959) in Table 1, p. 49. The correlations between "factors" or cluster primary axes, to two decimals, are given in Part II, Set B, of Table 2, p. 51, of Haggard et al (1959).

[a] The decimal points have been omitted in Cols. I to III inclusive in Sets B, E and F. The decimal is located two places to the left, i.e., -.74.

by some investigator as showing the "factor loadings" of 12 MMPI profiles on three axes defined by the clinical groupings of "hypertensive," "neurotic," and "psychotic." These data were computed from the 12 MMPI profiles as published by Haggard et al. The communalities were provided by E. A. Haggard. No "rotations" were made in obtaining these results; a single grouping of the profiles was made as specified by Haggard et al. (1959).

The Set B data are the orthogonal projections on the three centroid cluster vectors, a set of "primary axes," while the Set E values are the oblique or Cartesian projections on these same cluster vectors. The Set F data are the orthogonal projections on the reference axes or normals, each of which is orthogonal to all but one of the cluster vectors. (The sets of primary axes and reference axes are each collinear with the "inverse vectors" of the other set.) If another definition of the "factors" were used, as in an oblimax solution, three (or four) additional "factor matrices" might be computed. Comparisons between the results for different definitions of the factors could be made in terms of any of the corresponding (comparable) factor matrices.

The "communalities" or "common factor variances" of the 12 MMPI profiles are shown in the h^2 column of Table 1. Since three of these values exceed unity, the upper limit for "admissable" values of the communalities, the analysis is a "Heywood" case and not a "proper" factor analysis (Thurstone, 1947). The

presence of five residuals (computed from the Set B data of Table 1) exceeding .04 in absolute value and of the three communalities of greater than unity indicate that another factor probably should be defined; defined, however, only if a factor analysis is considered worth making with intercorrelations based on 9 observations (here, 9 MMPI scales).

Direct Correlation Methods

A purportedly new approach to the problem of determining profile groups is presented by Haggard et al. (1959) through the use of "empirical criterion profiles," i.e., the average profiles for a set of cases grouped in accordance with some set of rules, rules which may include nonprofile information (p. 51–52). The approach is hardly new since Burt in 1941 makes reference to his use of it in 1931 (Burt: 1937, 1941).

A more basic point arises from the discussion by these authors of the comparison, with reference to their Table 2, of their Set B data (the covariances between cluster factors and 12 MMPI profiles) with their Set C data (the correlations between three average criterion profiles and 12 MMPI profiles). They say that the methods agree closely, that the results are practically identical, but that they are not prepared to argue that either set is approximated by the other (p. 52). Apparently the authors are not aware of the fact that the results are not identical solely because of the defining procedures they happen to have used. The insertion of communalities rather than unities in the matrix of correlations for the Set B data in the first source of the small differences. The second source is their definition of average profiles in terms of the usual MMPI scaled scores rather than in terms of standard scores (over the 9 scales); this second source of differences introduces into their definitions differential weighting of the profiles in terms of unequal profile standard deviations. The "criterion profiles" cannot, as they suggest, be considered as "computed on equalized means and sigmas" (p. 54) since the profile standard deviations are not equal.

The equivalence of the "correlation of sums" methods and of the "sum of correlations" method for standard scores noted above as well as the ease of computation of correlations with totals or subtotals of scores is well known and has been discussed in these terms in some detail by Holzinger and Harman (1941) Holzinger (1944), and Richardson (1941) among others. Because of this equivalence, it is not correct to suggest that criterion profiles are more "intuitively meaningful" as a "definition of a factor" than are linear combinations of correlations (p. 53); the differences are only computational ones or are differences in definitions. This same basic point is also involved in the ambiguous statement dealing with the existence of a "general factor" in the footnote to Table 4, page 55. Furthermore, if any appreciable amount of "reassignment of variables to new groups," a problem mentioned by the authors (p. 55), were to arise, the flexibility available in the method using sums of weighted correlations or covariances for forming new groups might well overbalance the initial computational convenience of the "correlation of sums" method.

In presenting a second direct correlation method, that of the intraclass correlations, the authors emphasize the well-known transformation to standard scores imposed by the definition of a product moment correlation. As a result of this definition the product moment correlation

is limited to providing a definition of "profile shape"; one cannot combine with this measure of "shape" the independently defined profile attributes of "level" and "scatter." Several distinct composites of the three profile attributes, as pointed out by the authors, can be defined by an investigator using the intraclass coefficient. It seems worth noting, however, that Cronbach has recently raised a serious question as to the advisability of defining such "global" or composite concepts in his treatment of dyadic scores (1958).

Only two additional points relating to the presentation of the intraclass correlation will be made here. The first point is that differences in means and variances are not, as the authors state, like "poor Clementine," "lost and gone forever when r is used" (p. 53). The data are available and might even be used to test hypotheses regarding certain interactions as well as the homogeneity of the means when the study is properly conducted. (Lindquist, 1953).

The second point involves the stituation in which an investigator "may wish to form groups in terms of one or more a priori 'ideal' profiles which are based on theoretical considerations" (p. 55). It is simply not correct to state that "with the procedures discussed thus far, it is not possible to form groups around such a priori profiles" (p. 55). "Theoretical" profiles have been written and the profiles of individuals compared with "hypothetical types" or with "theoretical standard persons" for many years by Burt (1941) and others. Every possible comparison between any arbitrary sets of numbers and one or more profiles can also be made; in particular, observed profile "levels," "scatter," and "shape" indices can be compared to a priori values of these defined concepts. Correlations with such a priori "shape" values could be included in a factor analysis or "direct correlation" study as well as in an investigation using intraclass correlation coefficients. Whether any of these comparisons is empirically useful is another matter.

Concluding Remarks

A few of the more fundamental misconceptions or technically incorrect statements contained in the paper by Haggard et al. (1959) have been noted. Other technically questionable discussions include the authors' presentation of the orthogonal centroid method (p. 50), the possible assignment of cases to groups defined by factor analytic methods (p. 56), and the testing of statistical hypotheses using sets of related observations (p. 48, p. 57).

REFERENCES

Burt, C. Correlations between persons. *Brit. J. Psychol.*, 1937, **28**, 59–96.

Burt, C. *The factors of the mind.* New York: Macmillan, 1941.

Cronbach, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagirui and L. Petrullo (Eds.), *Person perception and interpersonal behavior.* Stanford Univer. Press, 1958, pp. 353–379.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between persons. *Psychol. Bull.*, 1953, **50**, 456–473.

Guttman, L. General theory and methods for matrix factoring. *Psychometrika*, 1944, **9**, 1–16.

Guttman, L. Multiple group methods for common factor analysis: Their basis, computation, and interpretation. *Psychometrika*, 1952, **17**, 209–222.

Haggard, E. A., Chapman, J. P., Isaacs, K. S., & Dickman, K. Intraclass correlation vs. factor analytic techniques for determining groups of profiles. *Psychol. Bull.*, 1959, **56**, 48–57.

HARMAN, H. H. The square root method and multiple group methods of factor analysis. *Psychometrika*, 1954, **19**, 39–55.

HOLZINGER, K. J. Factoring test scores and implications for the method of averages. *Psychometrika*, 1944, **9**, 155–167.

HOLZINGER, K. J., & HARMAN, H. H. *Factor analysis*. Chicago: Univer. Chicago Press, 1941.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. New York: Houghton Mifflin, 1953.

RICHARDSON, M. W. Combination of measures, supplementary study D. In P. Horst (Ed.), *Prediction of personal adjustment*. New York: Soc. Sci. Res. Coun. Bull., 1941, No. 48.

THURSTONE, L. L. *Multiple factor analysis*. Chicago: Univer. Chicago Press, 1947.

(Received April 15, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57 No. 2, 1960

REPLY TO PROFESSOR BECHTOLDT'S CRITIQUE

ERNEST A. HAGGARD
Department of Psychiatry, University of Illinois

After rereading Bechtoldt's critique of our paper (Haggard, et al., 1959),[1] I still find it difficult to understand fully why he became so exercised over it. True, many of the comments in his section on "Factor Analytic Techniques" are interesting and informative—and generally corrective—but they are also essentially irrelevant to our paper and its purpose.

It is obvious that Bechtoldt likes factor analysis and is steeped in it. It should also be obvious that I have a somewhat negative loading on the factor entitled "One should use factor analysis regardless of whether other methods can get the job done as well or better." There is clear difference of opinion as to whether, when, or what factor analytic techniques should be used in particular cases. Fortunately or unfortunately, I have never been enamoured of this group of techniques, and if I were to deal with problems of the sort usually associated with these methods, I would prefer an approach to data analysis of the type proposed by Creasy (1957).[2]

Some of Bechtoldt's remarks in his sections entitled "Direct Correlation Methods" and "Concluding Remarks" call for more specific comment, primarily because I think he misses the point from time to time. For example, his apparent reference to the discussion under the first italicized heading on our p. 54, namely, "When profile means and sigmas are equalized" has to do with the simple fact that for product-moment r's (over the nine subscales or profiles for any pair of *S*s), mean = 0 and sigma = 1 by definition. This is just the point which he emphasizes a couple of paragraphs later when he speaks of "the well-known transformation to standard scores imposed by the definition of a product moment correlation." Also, it may be that the expression "poor Clementine" is unfortunate stylistically, but I still fail to see how the coefficient r—by itself—provides information as to the means and sigmas of the variates correlated. Of course the means and sigmas are involved in the computation of r but, even though they may be of great use, they are all too often ignored in practice. Along these same lines, Bechtoldt should have gone much further than he did at the end of this paragraph: he should have emphasized along with Tukey (1951), Lindquist (1953), myself (1958) and

[1] As senior author, I take full responsibility for any "erroneous," "misleading," "misconceived," or "ambiguous" statements in the paper criticized by Bechtoldt, and so am replying to his comments.

[2] In a thoughtful and sobering discussion published several years ago, Horst (1950) outlined some of the well-known limitations of factor analysis, such as the large amount of data needed, the time required to perform the computations, and a number of technical difficulties. With reference to the first point, Horst stated that "If you want to come out with results of any consequence you should have 50 or 60 variables or tests and at least 500 cases. . . . I would not be inclined to take very seriously the results of any factor analysis involving psychological tests, which falls far short of 10,000 man hours of testing time" (p. 53). Creasy (1957) also mentioned various limitations of factor analytic techniques, as have so many other workers in this area. But these matters, which are quite relevant to Bechtoldt's remarks, are too numerous and complex to discuss here.

many others that the estimation of components of variance is a much more useful approach to answering many research questions than r is or ever can be.

In his next paragraph, Bechtoldt has us on the ropes when he cites the statement "with the procedures discussed thus far, it is not possible to form groups around such a priori profiles." We should have said "not practicable" (instead of using the too-strong term "not possible") to indicate that, for the majority of research workers, these procedures are not possible from a practical point of view. Finally, in his concluding paragraph, he appears to chide us (following his phrase "other technically questionable discussions") for possibly suggesting "the testing of statistical hypotheses using sets of related observations." Now, I thought we were clear on that point at least. We took pains to observe that, although multivariate data can be analyzed properly only by the appropriate multivariate statistical techniques, approximate produres for pattern analytic studies are available which utilize most of the information in the data without violating certain important statistical assumptions.

In conclusion, I wish to thank Professor Bechtoldt for bothering to rework our data and for presenting his Table 1—which illustrates so nicely the very point we made on our p. 51, namely, that "any decision based on the italicized values in Table 2 would result in the same choice as to which profiles belong together to form the three groups." According to his results, the *S*s would be grouped exactly as they are in our Table 2. And I will leave it to other readers, particularly those who work with small groups of *S*s and who do not have access to computer facilities, to decide which methods may be more useful for their purposes.

REFERENCES

CREASY, MONICA A. Analysis of variance as an alternative to factor analysis. *J. Roy. Statist. Soc., Series B (Methodological)*, 1957, **19**, 318–325.

HAGGARD, E. A. *Intraclass correlation and the analysis of variance*. New York: Dryden, 1958.

HAGGARD, E. A., CHAPMAN, JEAN P., ISAACS, K. S., & DICKMAN, K. Intraclass correlation vs. factor analytic techniques for determining groups of profiles. *Psychol. Bull.*, 1959, **56**, 48–57.

HORST, P. Uses and limitations of factor analysis in psychological research. *Proc., 1949 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1950, 50–56.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

TUKEY, J. W. Components in regression. *Biometrics*, 1951, **7**, 33–70.

(Received April 27, 1959)

PSYCHOLOGICAL BULLETIN
Vol. 57, No. 2, 1960

A REVIEW OF HEARING IN AMPHIBIANS AND REPTILES[1]

THOMAS E. McGILL[2]
Williams College

In the course of the evolutionary development of the vertebrate ear important changes took place as animals became terrestrial. The aquatic ear of the fishes was next to useless without the development of specialized accessory structures for the purpose of matching the impedance between aerially conducted sounds and the fluids of the inner ear. The middle ear mechanism of man is such an impedance-matching device. In the lower vertebrates the function is served by one rod-like bone, called the columella, which extends from the tympanic membrane to the inner ear. This structure is typical in both amphibians and reptiles.

Knowledge of the structure of the auditory mechanism in these classes far exceeds knowledge of the function. Disagreement exists on both theoretical and empirical grounds as to whether the hearing organs of certain modern amphibians and reptiles are vestigial or rudimentary. If they are vestigial, then, presumably, the animals concerned are deaf; if they are rudimentary, the animals should have some sense of hearing.

Presumptive evidence of hearing is obtainable from electrophysiological studies. Positive results from such studies demonstrate functional activity of the auditory system from the drum membrane to that part of the system from which the recording was taken. However, proof of hearing can only result from some behavioral manifestation on the part of the animals to sounds. These behavioral manifestations are of two general types: some natural reaction of the animals to sound, or a trained reaction. Following is a review of studies concerned with hearing in two orders of the class Amphibia and three orders of the class Reptilia.

Amphibia

Order Urodela—salamanders. Ferhat-Aket (1938) was able to demonstrate that *Amblystoma mexicanum* and other urodele species could hear within the frequency limits of 32 cps to 244 cps. Head raising, snapping, and restless movements were conditioned to the sounds of Edelmann whistles, mechanically actuated tuning forks, organ pipes, and a cello when these stimuli were followed by food. Kuroda (1926), using breathing rate changes to acoustic stimulation, could find no evidence of hearing in six newts. His stimuli were hand claps, whistles, tuning forks, electric bells, pistol shots, a harmonica, and an Edelmann whistle.

Order Salientia (Anura)—frogs and toads. Yerkes (1905) has presented evidence of hearing in frogs. He found that whistles and bells decreased the rate of respiration; he noted that stroking one frog and causing him to croak would result in other frogs croaking; he found that the reaction time to visual or tactual stimuli could be reduced if an auditory stimulus had been sounded within one second before the visual or tactual stimulation. The latter procedure resulted in an estimate of an effective frequency range of 25 cps to 5,000 cps. Bruyn

[1] This research was supported by Grant G-6119 from the National Science Foundation.

[2] At present, United States Public Health Service Postdoctoral Fellow, University of California, Berkeley, California.

and Van Nifterik (1920) discovered that a sound signal such as Yerkes used with frogs would also reduce reaction time to tactual stimuli in the toad. The sound signal in the toad was effective for periods up to 10 seconds.

Bajandurow and Pegel (1932) reported that breathing changes and jumping were readily obtained in frogs to "whistle-tones" when shock was used as the unconditioned stimulus. The responses were unstable and *Ss* had to be retrained each day. Corbeille (1929) used natural breathing rate changes in frogs as an index of hearing for the sounds produced by a Cambridge vibrator. He estimated the effective frequency range to be 100 cps to 8000 cps. Kuroda (1926) also used natural breathing rate as an index and found "some evidence for hearing in adult frogs and toads."

The upper frequency limit reported by Yerkes (1905) and Corbeille (1929) is surprising in view of the results of a recent study by Strother (in press). Strother at first made an unsuccessful attempt to condition bullfrogs to pure tones using shock and leg-flexion. He then operated upon the animals and recorded electrophysiological responses from the inner ear. The upper limit of response was found to be 3000 cps and the response at any frequency or intensity was of very low magnitude. It is possible that species differences account for the differing results.

Reptilia

Order Crocodilia—crocodiles and alligators. Beach (1944) reported that one captive alligator roared consistently when a 57 cps tone was presented. Other behavioral manifestations of hearing were evident up to 341 cps. Three smaller animals did not roar but showed evidence of hearing by turning to the sound source or by snapping. In reference to the problems involved in the care of reptiles, Pope (1950, p. 327) wrote, "In contrast to most other reptiles, crocodilians will learn to come when called at meal times," which, of course, implies that they can hear.

Adrian (1938) and Adrian, Craik, and Sturdy (1938) successfully recorded both "cochlear" potentials and action potentials from alligators. Wever and Vernon (1957), using the spectacled caiman (*Caiman sclerops*), recorded electrical responses from the inner ear for tones in the frequency range of 200 cps to 6000 cps. The frequencies to which the animals were most sensitive were located between 100 cps and 3000 cps.

Order Chelonia—turtles and tortoises. Concerning hearing in turtles Munn (1955, p. 97) has written, "Snakes, turtles, and similar vertebrates are believed to be entirely deaf. No experiment on these animals has given the least evidence that they can even hear noise." Actually, there are two rather obscure studies which profess to demonstrate hearing in the turtle. Andrews (1915) was able to train *Chrysemys* to distinguish between the sound of a whistle and the sound of a bell when one was made the positive signal for feeding. Poliakov (1930) established conditioned head withdrawal to a variety of tones and noises in *Emys orbicularis*. Kuroda (1923, 1925, 1926) on the other hand, failed to confirm Andrews' findings.

Recently, electrophysiological studies dealing with turtles have been carried out by Wever and Vernon (1956a, 1956b, 1956c). They found that the turtle's ear was uniformly highly sensitive for faint tones in the region from 100 cps to about 700

cps. The potentials then fell off rapidly up to 3000 cps, beyond which point injurious intensities were required to produce a measurable response.

Order Squamata—snakes and lizards. Concerning the evolution of the snake, Huxley (1953, p. 67) wrote " . . . all the circumstantial evidence makes it reasonably certain that the ancestors of the group had to pass through a stage of existance underground as deaf, half-blind, and legless burrowing lizards." After re-emerging, the line ". . . reacquired much of its power of vision (but not of hearing) and achieved new evolutionary success as snakes." Kuroda (1926) failed to find any behavioral evidence for hearing, and Adrian (1938) could record no response from the eighth cranial nerve in the grass snake to sounds; however, the nerve did respond to tactile stimulation. Wever and Vernon have recently carried out investigations of the electrophysiological responses of the inner ear of several species of snakes.[3] The resulting potentials were found to resemble those of turtles in terms of range, and to be only somewhat reduced in terms of magnitude.

The snake's close relative, the lizard, definitely demonstrates hearing according to Kuroda (1926). Kuroda found that lizards would open their eyes to tonal stimulation. He likened this reflex to Preyer's reflex "as a reliable clue to make sure objectively of the normal state of audition." Using a Galton whistle he established the upper frequency limit at 9675 vs per second (4837.5 cps).

[3] E. G. Wever and J. A. Vernon. Personal communication. 1959.

Summary

If one is willing to accept the behavioral evidence available, then adult amphibians—salamanders, frogs and toads—can hear. Among the reptiles, alligators and lizards can hear, snakes cannot, and the hearing of turtles is in question. The electrophysiological evidence indicates functional activity of the auditory system up to, and including, the sensory cells in every species tested. Because there has not been a single behavioral study successfully carried out since the development of modern electronic devices for the production and control of sound, we know nothing of the animals' ability to hear pure tones, nor of their absolute limens in terms of intensity, nor is there any certainty of the frequency range for any species. The present state of knowledge of hearing in amphibians and reptiles is not commensurate with the importance of these classes in the study of the evolution of the sense of hearing.

REFERENCES

ADRIAN, E. D. The effect of sound on the ear in reptiles. *J. Physiol.*, 1938, **92**, 9–11.

ADRIAN, E. D., CRAIK, K. J. W., & STURDY, R. S. The electrical responses of the auditory mechanism in coldblooded vertebrates. *Proc. Roy. Soc.*, 1938, **125**, 435–455.

ANDREWS, O. The ability of turtles to discriminate between sounds. *Bull. Wisconsin Nat. Hist. Soc.*, 1915, **13**, 189–195.

BAJANDUROW, B. I., & PEGEL, W. A. Der bedingte Reflex bei Fröschen. *Zsch. Physiol.*, 1932, **18**, 284–297. (*Psychol. Abstr.*, **8**:3019).

BEACH, F. A. Responses of captive alligators to auditory stimulation. *Amer. Nat.*, 1944, **78**, 481–505.

BRUYN, E. M. M., & VAN NIFTERIK, C. G. M. Influence du son sur le reaction d'une excitation tactile chez les grenouilles et les crapauds. *Arch. néerl. Physiol. Hom. Anim.*, 1920, **5**, 363–379.

CORBEILLE, C. L'influence des vibrations

acoustiques sur la respiration chez la grenouille et certains mammifères. *C. R. Soc. biol.*, 1929, **101**, 113–115.

FERHAT-AKET, S. Untersuchungen über den Gehorsinn der Amphiben. *Z. vergl. Physiol.*, 1938, **26**, 253–281.

HUXLEY, J. S. *Evolution in action*. New York: Harper, 1953.

KURODA, R. Studies of audition in reptiles. *J. comp. Psychol.*, 1923, **3**, 27–36.

KURODA, R. A contribution to the subject of the hearing of tortoises. *J. comp. Psychol.*, 1925, **5**, 285–291.

KURODA, R. Experimental researches on the sense of hearing in lower vertebrates, including reptiles, amphibians, and fishes. *Comp. Psychol. Monog.*, 1926, **3**(16) 1–50.

MUNN, N. L. *The evolution and growth of human behavior*. New York: Houghton Mifflin, 1955.

POLIAKOV, K. Zur Physiologie des Riech— und Horanalysators bei der Schildkröte *Emys orbicularis*. *Russ. Fiziol Zh.*, 1930, **13**, 161–178.

POPE, C. H. In E. J. Farris (Ed.), *The care and breeding of laboratory animals*. New York: Wiley, 1950.

STROTHER, W. F. The electrical response of the auditory mechanism in the bullfrog *Rana catesbeiana*. *J. comp. physiol. Psychol.*, in press.

WEVER, E. G., & VERNON, J. A. Auditory responses in the common box turtle. *Proc. Natl. Acad. Sci.*, 1956, **42**, 962–965. (a)

WEVER, E. G., & VERNON, J. A. The sensitivity of the turtle's ear as shown by its electrical potentials. *Proc. Nat. Acad. Sci.*, 1956, **42**, 213–220. (b)

WEVER, E. G., & VERNON, J. A. Sound transmission in the turtle's ear. *Proc. Nat. Acad. Sci.*, 1956, **42**, 292–299. (c)

WEVER, E. G., & VERNON, J. A. Auditory responses in the spectacled caiman. *J. cell. comp. Physiol.*, 1957, **50**, 333–341.

YERKES, R. M. The sense of hearing in frogs. *J. comp. neurol. Psychol.*, 1905, **15**, 279–304.

(Received April 28, 1959)

A GLOSSARY

OF SOME TERMS USED IN THE OBJECTIVE SCIENCE OF BEHAVIOR

By William S. Verplanck

Provides an empirical vocabulary in the science of human and animal behavior

Familiarizes readers with developments in the study of animal behavior

Clarifies concepts used by behaviorists and ethologists

Price $1.00

Order from:

American Psychological Association
1333 Sixteenth St., N.W.
Washington 6, D. C.

GEORGE BANTA COMPANY, INC., MENASHA, WISCONSIN, U.S.A.